JAZYKOVEDNÝ ČASOPIS

JAZYKOVEDNÝ ÚSTAV ĽUDOVÍTA ŠTÚRA

SLOVENSKEJ AKADÉMIE VIED

2

ROČNÍK 72, 2021

sciendo

SAP

JAZYKOVEDNÝ ČASOPIS
VEDECKÝ ČASOPIS PRE OTÁZKY TEÓRIE JAZYKA

JOURNAL OF LINGUISTICS
SCIENTIFIC JOURNAL FOR THE THEORY OF LANGUAGE

Vydáva/Published by: Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied
- v tlačenej podobe vo vydavateľstve SAP – Slovak Academic Press, s. r. o.
- elektronicky vo vydavateľstve Sciendo – De Gruyter
  https://content.sciendo.com/view/journals/jazcas/jazcas-overview.xml

Adresa redakcie/Editorial address: Jazykovedný ústav Ľ. Štúra SAV, Panská 26, 811 01 Bratislava
Kontakt: gabriela.mucskova@juls.savba.sk
Elektronická verzia časopisu je dostupná na internetovej adrese/The electronic version of the journal is available at: http://www.juls.savba.sk/ediela/jc/

Vychádza trikrát ročne/Published triannually

Dátum vydania aktuálneho čísla (2021/72/2) – október 2021

CiteScore 2020: 0,4
SCImago Journal Rank (SJR) 2020: 0,186
Source Normalized Impact per Paper (SNIP) 2020: 0,876

JAZYKOVEDNÝ ČASOPIS je evidovaný v databázach/JOURNAL OF LINGUISTICS is covered by the following services: Baidu Scholar; Cabell's Directory; CEJSH (The Central European Journal of Social Sciences and Humanities); CEEOL (Central and Eastern European Online Library); CNKI Scholar (China National Knowledge Infrastructure); CNPIEC – cnpLINKer; Dimensions; DOAJ (Directory of Open Access Journals); EBSCO (relevant databases); EBSCO Discovery Service; ERIH PLUS (European Reference Index for the Humanities and Social Sciences); Genamics JournalSeek; Google Scholar; IBR (International Bibliography of Reviews of Scholarly Literature in the Humanities and Social Sciences); IBZ (International Bibliography of Periodical Literature in the Humanities and Social Sciences); International Medieval Bibliography; J-Gate; JournalGuide; JournalTOCs; KESLI-NDSL (Korean National Discovery for Science Leaders); Linguistic Bibliography; Linguistics Abstracts Online; Microsoft Academic; MLA International Bibliography; MyScienceWork; Naver Academic; Naviga (Softweco); Primo Central (ExLibris); ProQuest (relevant databases); Publons; QOAM (Quality Open Access Market); ReadCube; SCImago (SJR); SCOPUS; Semantic Scholar; Sherpa/RoMEO; Summon (ProQuest); TDNet; Ulrich's Periodicals Directory/ulrichsweb; WanFang Data; WorldCat (OCLC).

ISSN 0021-5597 (tlačená verzia/print) **ISSN 1338-4287 (verzia online)** **MIČ 49263**

JAZYKOVEDNÝ ČASOPIS

JAZYKOVEDNÝ ÚSTAV ĽUDOVÍTA ŠTÚRA SLOVENSKEJ AKADÉMIE VIED

2

ROČNÍK 72, 2021

NLP, Corpus Linguistics and Interdisciplinarity

SLOVKO 2021

Tematické číslo Jazykovedného časopisu venované počítačovému spracovaniu prirodzeného jazyka, korpusovej lingvistike a interdisciplinárnemu výskumu.

Prizvaní editori:
Kristína Bobeková
Lucia Gibľáková
Jana Levická
Miroslav Zumrík

sciendo

CONTENTS

INTERDISCIPLINARY RESEARCH BASED ON CORPORA

FOREWORD

The reaching out of Corpus Linguistics to other linguistic and scientific disciplines is not a matter of fashion, but a natural phenomenon. The interdisciplinary dimension of Corpus Linguistics follows naturally from the fact that corpus linguistics – as stated by editors of the book "The Corpus Linguistics Discourse" (2018) Anna Čermáková and Michaela Mahlberg – "can make connections across linguistic disciplines that do not easily seem to get together" (p. 6). Among the strengths of Corpus Linguistics, the editors also mention "focus on the identification of tendencies and patterns of mainstream language use" and they continue: "The more repetitions we find of patterns and meanings the clearer the picture becomes." Contributions to the **11th international biannual linguistic conference SLOVKO 2021**, which has been given the title *NLP, Corpus Linguistics and Interdisciplinarity*, that are presented in this special issue of the *Journal of Linguistics* (Jazykovedný časopis), together sketch a hopefully similar picture of the Corpus Linguistic discourse.

The picture is definitely varied, that is, thematically diverse, as the authors aim at a wide array of linguistic phenomena: artificial homonyms, verb valency, word order, linguistic prototypes, adjectival constructions and others. The contributors work within various language families (Slavic, Germanic, Asian) and practically all major linguistic levels (morphological, lexical, syntactical, phrasal or semantic), while they also focus on several genres (19th century essays, biblical texts), and make use of multiple linguistic approaches to the interaction between meaning and form (systematic, historical, cognitive linguistics, to name just a few), and, very importantly, operate at the interdisciplinary intersection with (social) media and (specialized) discourse studies or translatology. From this respect, our aim as the editors has been to offer a representative overview over the many research questions and projects conducted in today's linguistics in both Slavic (Slovakia, Czech Republic, Poland, Slovenia, Croatia, Russia, Ukraine, Bulgaria) and Germanic countries (Austria, Germany and Sweden), thus also creating a platform for an international scientific dialogue.

The impressively "colorful" picture of today's Corpus Linguistics needs, however, a certain unifying frame, so that the variety can be perceived and appreciated as such. A thematic divergence is, in other words, helped by a synthesizing counterweight – a methodological convergence. In this respect, our aim has also been to show that the varied contributions share to a certain degree a common denominator: be it an interest in both corpus-based and -driven empiric

and systemic inquiry into written and spoken language (the part *Corpus-Based/ Corpus-Driven Linguistic Research*), an interest in technological development of annotation and visualization tools, creating pedagogical resources and applications for processing of language data (the part *Corpus Building and Natural Language Processing*) or in interdisciplinary connection between conceptual and technological instruments of linguistics with the research in other domains (the part *Interdisciplinary Research Based on Corpora*). All of the aforementioned linguistic and computational task are in their own right able to help with processing raw linguistic material that can be studied and used by both researchers, language users in general and language learners in particular. It is therefore our hope that the contributions presented in the issue will not draw only a simple picture with just one red thread, but rather picture of many threads, bringing about interdisciplinary connections between linguistic and other scientific domains that are set within an ever-growing digital and media environment of nowadays's world and to a great extent enabled by the very unprecedented potential of corpus resources, tools and methods.

Editors

PREDHOVOR

Presah korpusovej lingvistiky do iných lingvistických a vedeckých disciplín nie je módnou záležitosťou, ale prirodzeným javom. Interdisciplinárny rozmer korpusovej lingvistiky pritom prirodzene vyplýva z faktu, že korpusová lingvistika – ako tvrdia editorky publikácie „Diskurz korpusovej lingvistiky" (The Corpus Linguistics Discourse, 2018) Anna Čermáková a Michaela Mahlbergová – „dokáže prepájať jazykovedné disciplíny, ktoré k sebe zdanlivo majú navzájom ďaleko" (s. 6). Editorky publikácie tiež ako silnú stránku korpusovej lingvistiky uvádzajú jej „zameranie na identifikáciu tendencií a pravidelností v bežnom používaní jazyka", a pokračujú: „Čím frekventovanejšie pravidelnosti a významy odhalíme, tým jasnejší obraz získame." Príspevky **11. medzinárodnej jazykovednej konferencie SLOVKO 2021** s názvom *Počítačové spracovanie prirodzeného jazyka, korpusová lingvistika a interdisciplinárny výskum*, ktoré prináša toto tematické číslo Jazykovedného časopisu, spoločne načrtávajú azda podobne jasný obraz korpusovolingvistického diskurzu.

Tento obraz sa rozhodne vyznačuje diverzitou, to znamená tematickou rôznorodosťou, keďže sa autori jednotlivých príspevkov zameriavajú na široké spektrum jazykových javov: homonymá, valenciu slovies, slovosled, jazykové prototypy, adjektívne konštrukcie a iné. Prispievatelia sa pohybujú v rámci rozličných jazykových rodín (slovanskej, germánskej, ázijskej) a prakticky všetkých hlavných jazykových úrovní (morfologickej, lexikálnej, syntaktickej, frazeologickej či sémantickej), pričom sa zaoberajú rôznymi žánrami (esejami z devätnásteho storočia, biblickými textami) a využívajú niekoľko lingvistických prístupov k skúmaniu interakcie medzi významom a formou (za všetky napríklad prístupy systematickej, historickej či kognitívnej lingvistiky) a, čo je veľmi dôležité, uskutočňujú výskum na interdisciplinárnej križovatke so štúdiami (sociálnych) médií a (špecializovaných) diskurzov či translatológie. Z tohto hľadiska sme si ako editori kládli za cieľ ponúknuť reprezentatívny prehľad mnohorakých výskumných otázok a projektov uskutočňovaných na poli súčasnej lingvistiky v slovanských (Slovensko, Česká republika, Slovinsko, Poľsko, Chorvátsko, Rusko, Ukrajina, Bulharsko) i germánskych krajinách (Rakúsko, Nemecko a Švédsko), čím sme tiež chceli vytvoriť podmienky pre širší medzinárodný vedecký dialóg.

Pôsobivo „pestrý" obraz korpusovej lingvistiky si však vyžaduje istý jednotiaci rámec, aby sa rôznorodosť ako taká mohla vnímať a oceniť. Tematickej divergencii teda inými slovami napomáha syntetizujúca protiváha – metodologická konvergencia. Z tohto hľadiska sme sa zasa usilovali ukázať, že rôznorodé príspevky do istej miery zdieľajú spoločného menovateľa: či už záujem o korpusmi podporovaný alebo riadený empirický i systematický výskum písanej a hovorenej reči (časť *Korpusovo založený a korpusovo riadený lingvistický výskum*), záujem o technologický rozvoj anotačných a vizualizačných nástrojov, o tvorbu pedagogických zdrojov a aplikácií na spracovanie jazykových dát (časť *Budovanie korpusov a počítačové spracovanie*

prirodzeného jazyka), a zároveň záujem o interdisciplinárne prepájanie pojmového a technologického inštrumentária jazykovedy s výskum v ďalších oblastiach (časť *Interdisciplinárny výskum založený na korpusoch*). Všetky vyššie uvedené úlohy jazykovedného a počítačového výskumu dokážu svojím spôsobom napomôcť pri spracúvaní jazykového materiálu, ktorý môžu študovať a využívať vedci, používatelia jazyka vo všeobecnosti a ľudia pri štúdiu jazykov zvlášť. Dúfame preto, že príspevky v tomto čísle nenačrtnú len prostý obraz s jednou červenou niťou, ale s viacerými niťami, a že sa tak ukážu interdisciplinárne prepojenia medzi lingvistickými a ostatnými vedeckými oblasťami. Tieto prepojenia sú zasadené do neustále rastúceho digitálneho a mediálneho prostredia dnešného sveta a do veľkej miery ich umožnil práve bezprecedentný potenciál korpusových zdrojov, nástrojov a metód.

Editori

CORPUS-BASED/CORPUS-DRIVEN LINGUISTIC RESEARCH

sciendo
DOI 10.2478/jazcas-2021-0030

FROM GRAPHEMATICS TO PHRASAL, SENTENTIAL, AND TEXTUAL SEMANTICS THROUGH MORPHOSYNTAX BY MEANS OF CORPUS-DRIVEN GRAMMAR AND ONTOLOGY: A CASE STUDY ON ONE TIBETAN TEXT

ALEKSEI DOBROV[1] – MARIA SMIRNOVA[2]
[1] LLC "AIIRE", Saint Petersburg, Russia
[2] Saint Petersburg State University, Saint Petersburg, Russia

DOBROV, Aleksei – SMIRNOVA, Maria: From graphematics to phrasal, sentential, and textual semantics through morphosyntax by means of corpus-driven grammar and ontology: A case study on one Tibetan text. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 319 – 329.

Abstract: This article presents the current results of an ongoing study of the possibilities of fine-tuning automatic morphosyntactic and semantic annotation by means of improving the underlying formal grammar and ontology on the example of one Tibetan text. The ultimate purpose of work at this stage was to improve linguistic software developed for natural-language processing and understanding in order to achieve complete annotation of a specific text and such state of the formal model, in which all linguistic phenomena observed in the text would be explained. This purpose includes the following tasks: analysis of error cases in annotation of the text from the corpus; eliminating these errors in automatic annotation; development of formal grammar and updating of dictionaries. Along with the morpho-syntactic analysis, the current approach involves simultaneous semantic analysis as well. The article describes semantic annotation of the corpus, required by grammar revision and development, which was made with the use of computer ontology. The work is carried out with one of the corpus texts – a grammatical poetic treatise *Sum-cu-pa* (VII c.).

Keywords: Tibetan language, computer ontology, Tibetan corpus, natural language processing, corpus linguistics, parsing

1 INTRODUCTION

This article discusses the development of a formal model (a grammar and a linguistic ontology) of the Tibetan language, including morphosyntax, syntax of phrases and super-phrasal units, and semantics that can perform the morpho-syntactic, syntactic, and semantic analysis. The engine is based on a consistent formal model of Tibetan vocabulary, grammar, and ontology, verified by and developed on the basis of a representative and manually tested corpus of texts, which includes the Basic Corpus of the Tibetan Classical Language [1] and the Corpus of Indigenous Tibetan Grammar Treatises [2], comprising 34,000 and 48,000 tokens, respectively [3]. Among the texts of our corpus, there are both prose and poetic texts.

Tibetan can reasonably be considered as one of the less-resourced languages. Despite the fact that scholars in different countries (Germany, United Kingdom, China, USA, Japan) are working on the tools for processing Tibetan texts, there is still no conventional standard for annotating a corpus of Tibetan language material. A number of recent studies were primarily aimed at developing solutions for such stages of Tibetan NLP as word segmentation and part-of-speech tagging. Some researchers use corpus methods to solve specific applied problems, as well as tasks in the field of history, literature, linguistics, and anthropology (e.g., [4], [5], [6]). Syntactic and semantic research of Tibetan has been comparatively weak.

The common problem of formal grammar development for less resourced languages is that in order to create an adequate formal model, a representative corpus of texts with reliable annotation must be created first, but in order to annotate a corpus, a formal model must already exist to form the basis of annotation. Thus, when working with the Tibetan language, we decided to implement an approach that allows us to gradually improve the formal grammar and develop the existing corpus. Initially, work was carried out simultaneously on all texts of the corpus. Various types of errors indicating typical problems (morphosyntactic ambiguity or lack of syntactic or semantic annotation) were analyzed and resolved taking into account all cases in the corpus representing a particular problem. When the problems became specific, we decided to analyze separate texts, sequentially improving linguistic software. The article describes the methods of working with the first text of the corpus – the Tibetan grammatical treatise *Sum-cu-pa* (VII c.) – the corrections required and the problems encountered. The choice of a poetic text is explained only by the fact that it is the shortest text in the corpus, consisting of 1356 tokens. It is convenient to use it to demonstrate the results of our work.

2 THE SOFTWARE TOOLS

This study is carried out within the framework of the AIIRE project [7] and with use of the technologies and tools of this project. AIIRE is a free open-source natural language understanding system, which is developed and distributed under the terms of GNU General Public License. This system implements the full-scale procedure of natural language understanding, from graphematics, through morphological annotation and syntactic parsing, up to semantic analysis.

2.1 Tokenization and morphological annotation

The module developed for the Tibetan language is designed taking into account the fact that since there are no separators between words in Tibetan writing, while morphology and syntax are significantly intermixed, the minimal (atomic) units of modeling (so-called atoms) are morphemes and their allomorphs, not words and their forms. Input string segmentation into such units (tokenization) cannot be

performed with standard tokenization algorithms, and is therefore performed in AIIRE by means of the Aho-Corasick algorithm (developed by Alfred V. Aho and Margaret J. Corasick [8]). This algorithm allows one to find all possible substrings of the input string according to a given dictionary. The algorithm builds a tree, describing a finite state machine with terminal nodes corresponding to completed character strings of elements (in this case, morphemes) from the input dictionary.

The Tibetan language module contains a dictionary of morphemes with their allomorphs, so that this tree can be created in advance at the build stage of the module and loaded as a shared library in the runtime, which minimizes its initialization time. The dictionary of morphemes contains grammatical and morphological attributes (grammemes) for each allomorph; these attributes are mapped onto classes of immediate constituents, so that the tree for the Aho-Corrasick algorithm contains just class and morpheme identifiers for each allomorph and doesn't need to store individual attributes. The module also contains a set of definitions that determines possible types of atoms (atomic units), possible attributes for each type of atom, possible values of each attribute, and restrictions on each attribute value.

Thus, AIIRE first builds all possible hypotheses of recognizing Tibetan atomic units in input strings, including overlapping substrings for separate hypotheses, and then brings them together immediately after they arise into trees of immediate constituents in all possible ways in accordance with the formal grammar, which models the Tibetan morphosyntax.

2.2 Syntactic parsing

The grammar is a combined grammar of immediate constituents and syntactic dependencies, which consists of the so-called classes of immediate constituents (CICs hereinafter). CICs are developed as python-language classes, with enabled built-in inheritance mechanism, and specify the following attributes: a semantic graph template which represents how the meaning of a constituent should be calculated from the meanings of its child constituents; lists of possible head and subordinate constituent classes; a dictionary of possible linear orders of the subordinate constituent in relation to the head and the meanings of each order; the possibility of head or subordinate constituent ellipsis; the possibility of non-idiomatic semantic interpretation [9, p. 146]. Currently, the formal grammar includes 507 CICs.

The formal grammar is developed in direct accordance with semantics, in a way that the meanings of syntactic and morphosyntactic constituents can be correctly calculated from the meanings of their child constituents in accordance with the Compositionality principle.

The Tibetan language module is integrated into AIIRE natural language processor, and the corpus texts are passed on for processing in unannotated form.

The results of linguistic processing are presented in the form of immediate constituent structures with semantic graphs, these structures forming the syntactic and semantic annotation of the corpus: the results of automatic text processing are loaded into the AIIRE corpus manager as the annotation of the corpus, upon which the corpus manager automatically searches for typical errors, indicating locations of incomplete annotation and possible inaccuracy. The four types of errors are: unrecognized units, combinatorial explosions, breaks in syntactic trees, and overlaps. Unrecognized fragments are those for which there are no syntactic trees in the annotation. Combinatorial explosions are cases of exponential growth in the number of parsing versions with respect to the length of the parsed text and, thus, the amount of its parsed ambiguous fragments increases. Breaks are positions in which the tree cannot be bound with any of its neighbours. Overlaps are fragments of text in which the syntactic trees overlap, not completely covering the text: a fragment covered by one tree includes the position of the beginning of the fragment covered by the next tree, but not the position at its end [10, p. 145].

This toolkit allows simultaneous work on the corpus annotation and on the improvement of the formal model behind this annotation, which is a new approach to the development of modules of the linguistic processor, ensuring continued verifiability of the formal model and its correspondence with the corpus material.

2.3 Semantic analysis

The ontology used for this research is a united, consistent classification of concepts that unite the meanings of linguistic units of the corpus texts, including morphemes and idiomatic morphemic complexes. To model a new concept, a researcher needs to create an expression entry in the ontology, and provide it with the meaning (translation) and description (or interpretation) in Russian.[1] The main source for establishing the basic meaning of each expression is a text or texts of the corpus where the expression is used. Regular use of the Tibetan explanatory dictionary helps to verify the choice of a Tibetologist, who edits the ontology. In some cases, translating and interpreting linguistic units (especially special terms) requires thematic dictionaries, thesauri and catalogues. In controversial cases, native speakers are involved.

The concepts are interconnected with different semantic relations. In addition to such semantic relations as synonymy, hyponymy, and hypernymy, the ontology models strictly specified relations between concepts such as the relation between a physical object and its parts (meronymy); between the agent and the actions that the agent can perform; between an action and objects towards which this action can

[1] The Russian language is the language of the software interface, including the ontology itself. In the ontology, Russian is also used for technical classes and to describe verbal semantics and relations between concepts.

be directed, etc. These relations allow semantic analysis of texts and lexical and syntactic disambiguation to be performed. The basic ontological editor is described with examples from the Tibetan ontology ([11], [12], [13]). As far as the authors of this article are currently aware, at the moment, AIIRE is the only system that actually implements not only word-sense, but also syntactic disambiguation by means of linguistic ontology without use of any statistical heuristics.

The ultimate goal of our project is to create a complete semantic annotation of all texts in the corpus. At the moment, 5230 concepts are modelled in the ontology, including the meanings of all lexical units (verbs, compounds and idiomatic expressions; 160 concepts in total) of the *Sum-cu-pa* grammar.

3 ANNOTATION DEVELOPMENT

The Thirty Verses (Tib. *Sum-cu-pa,* presumably 7th–9th centuries AD) is one of the first two Tibetan grammatical treatises that laid the basis for traditional Tibetan linguistics (Tib. *sgra'i rig-pa*). Tibetan proto-scientific texts have special structural features and methods of description, and use a large number of grammatical terms and special lexis. The characteristics of Tibetan poetic texts (omission of grammatical markers, ellipsis, adding syllables to comply with the poetic meter) also cause a number of difficulties in syntactic and semantic analysis and require updating of dictionaries and formal grammar development, along with the use of computer ontology.

3.1 New classes of atoms

Since the grammatical description in the *Sum-cu-pa* treatise begins with the structure of Tibetan syllables, after which the author describes the formation and meaning of various grammatical markers, it became necessary to create two new separate classes of atoms for letters and exponents of Tibetan morphemes and function words (e.g., the allomorph *gyi* of the morpheme *KYi* that expresses the genitive case meaning). The class Letter contains the letters of the Tibetan alphabet, and the class Exponent contains exponents of morphemes, which were used with metalinguistic meaning in the *Sum-cu-pa* grammar.

Letters and exponents act like nouns in the text – they can attach attributes (even to each other like in (1)); act as a subject or direct object of certain verbs. Separate classes of immediate constituents – entity argument (EntityArg[2]) and entity right argument (EntityRightArg) – were created for combinations "Letter/Exponent + intersyllabic delimiter" and "intersyllabic delimiter + Letter/Exponent". These classes, in turn, were embedded as arguments into transitive verbal phrases (TransitiveVP) and noun phrases with genitive (NPGen) respectively, which ensured correct syntactic parsing of sentences like (1).

[2] The names of the CICs from the formal grammar as they appear in syntactic graphs.

(1) སུ་ཡི་ཨུ་ཕྱིས་ནས། །དེ་ལ་གསུམ་པའི་དང་པོ་སྦྱར།
su yi u phyis nas / / de la gsu- pa ‘i dang-po sbyar
su GEN u remove-EL PDem DAT third GEN first join
‘After removing *u* [from the grammatical marker] *su*, add the first [letter] of the third [alphabet row] to it.’

All atoms that belong to the class Letter were assigned to the ontology concept *yi-ge* ‘grapheme’; while atoms that belong to the class Exponent inherit the concept ‘linguistic unit’. Thus, in order to avoid breaks in the annotation and ensure the correct semantic analysis of the genitive noun phrase from example (1) *su yi u* ‘*u* of [the marker] *su*’, the concept ‘linguistic unit’ was connected via the relation ‘to have a grapheme’, that is a subclass of the general genitive relation ‘to have any object or process’, with the concept *yi-ge* ‘grapheme’. The same concepts were used to limit the verb valencies.

3.2 Topicalized noun phrases

The *Sum-cu-pa* grammar is composed in heptasyllabic verses, united in *shlokas* (Sans. śloka, Tib. *sho-k+la*).[3] It is written in the most common meter that was used as a standard translation of Sanskrit *shlokas* and in much of the native poetry in classical Tibetan. The meter implies that every line should have seven syllables [14, p. 410]. In order not to violate the meter, the author of the treatise sometimes excessively uses the topicalizer *ni*. The text contains structurally identical phrases, in one of which there is a topicalizer, while in the other it is absent.

The topicalizer is used after the ordinary noun phrase in only nine out of twenty-four cases. In other cases, it is added excessively (also for filling out the meter) after function words denoting case meanings. For example, in (2) it is used after a noun phrase in the ablative.

(2) བདུན་པ་ལས་ནི་ཤ་མ་གཏོགས། །རྗེས་འཇུག་ཡི་གེ་བཅུ་རུ་འདོད། །
bdun pa las ni sha ma gtogs/ /rjes-’jug yi-ge bcu ru ‘dod/ /
seventh ABL TOP *sha* NEG-belong final_consonant phoneme ten TERM accept
‘As for [phonemes that are] from the seventh [row of the alphabet], all except the first letter belong to the final phonemes.’

For such cases the following CICs were created: topicalized noun phrase in ablative (TopicalizedAblativeNP); genitive (TopicalizedGenitiveNP); terminative (TopicalizedTerminativeNP); dative (TopicalizedDativeNP) and ergative (TopicalizedErgativeNP). In the ontology, the relation ‘to concern an object or process’ was created for topicalized noun phrases. This relation, in turn, was

[3] Element of a poetic text (analogue of a stanza). In the *Sum-cu-pa* treatise *shlokas* include from two to five seven-syllable lines.

connected via the relation ‘to have an object’ with the class of possible topics (that is, with any concept).

3.3 Zero nominalization

The term *zero nominalization* was suggested by N. Hill for morphologically finite forms occurring in syntactically nominal contexts [15, p. 5]. S. Beyer describes similar cases when the nominalizer *–pa* can be omitted between a tense stem of a verb and a bound role particle [14, p. 305]. Several examples of this phenomenon can be found in the *Sum-cu-pa* treatise. In most of them, the right context indicates that a verb functions as a noun. Usually, the nominalizer *–pa* occurs only after the last of several verbs, while the nominalization of the preceding ones is guaranteed by the choice of the conjunction particle *dang* like in (3), since *dang* occurs only after nouns or noun phrases ([14, p. 241], [15, p. 5]).

(3) འདྲི་[...]དང་སློག་[...]དང་བཤད་[...]རྣམས་ཀྱི། །མཚམས་སྦྱོར་སྒྲ་ལ་ཐོགས་མེད
‘dri [...] dang klog [...] dang bshad [...] rnams kyi/ /mtshams-sbyor-sgra la thogs med
ask CONJ read CONJ speak-PL GEN conjoining_marker DAT obstruct not_exist
‘There will be no difficulties with markers linking [words in the process of] writing, reading and explaining.’

In the first two cases of zero nominalization in (3), the choice of conjunction particle *dang* guarantees the interpretation of *‘dri* and *klog* as nominal forms. After *bshad*, we meet the plural marker *rnams* that also follows only nouns or noun phrases. The CIC poetic verbal noun (PoeticVN) was created for such cases in the formal grammar. This class was embedded in the CIC for noun phrases in the plural (InstanceNPPlural) and homogeneous noun phrases (InstanceNPGroup).

In examples (4) and (5), noun coordinators are used only once at the end of the passage.

(4) རྗེས་འཇུག་བཅུ་ཡི་སྦྱོར་བ་ནི། །མཉན་བསམ་བསྟན་པའི་དོན་དུ་སྦྱར། །
rjes-’jug bcu yi sbyor-ba ni/ /mnyan bsam bstan-pa’i don du sbyar/ /
final_consonant ten GEN join-NMLZ TOP listen think teach-NMLZ
‘As for adding of the ten final consonants, [these consonants] are added for listening, thinking and teaching.’

(5) སྡེབ་སྦྱོར་ལེགས་མཛད་མཁས་རྣམས
sdeb-sbyor legs mdzad mkhas rnams
poetry be_good do be_skilled-PL
‘[those who are] skilled in making good poetry’

In example (4), the nominalizer *–pa* is used once after three verbs – *mnyan* ‘to listen’, *bsam* ‘to think’, and *bstan* ‘to teach’ – that can be considered as homogeneous verbal phase. As this not a typical grammatical phenomenon for the Tibetan language, the special class PoeticHomogenVP was created and embedded into classes for verbal nominalization.

In example (5), we actually see five verbs with obviously different subordinate syntactic relations, but without any grammatical markers between them. Only the last verb takes the plural marker and thus can be undoubtedly treated as a case of zero nominalization. Still, this passage can be read in several ways. Disambiguation in this case will be discussed below (see section 3.5).

3.4 Equative verb omission

The equative verb *yin* expresses equation or identification of two patient participants (nouns or noun phrases of different length and complexity) [14, p. 255]. The *Sum-cu-pa* treatise demonstrates several omissions of the equative verb *yin* before the statement final particle *–O*. In most cases, the verb is omitted in a compound nominal predicate consisting of numeral like in (6).

(6) ཀཱ་ལི་སུམ་ཅུ་ཐམ་པའོ།
kA-li sum-cu tham-pa ‘o
consonant thirty even FIN
‘Consonants [are in the amount] thirty even.’

In the formal grammar, there was already a class for the copula group, consisting of an equative verb and a noun phrase. For cases like (6), the CIC for copula group with elliptic verb (EllCopulaGroup) was created, in which quantitative noun phrases were embedded.

3.5 Compounds with complex structure

Modeling of Tibetan compounds’ meanings was the first ontological task, since most combinatorial explosions contained compounds. As a result of this work, a classification of Tibetan verbal and nominal compounds was created. These types and their formal grammatical and ontological modeling are described in detail in [9].

In some cases, one of the components of a compound is itself a compound. In the *Sum-cu-pa* treatise, even more complex structures were discovered. In example (7), we found five verbal roots following each other without any markers between them. Relying on the context and several most authoritative commentaries on the *Sum-cu-pa* grammar, this passage can be read in the following way:

(7) སྡེབ་པར་སྦྱོར་བ་ལེགས་པར་མཛད་པའི་མཁས་པ།
sdeb [pa r] sbyor-[ba] legs-[pa r] mdzad-[pa ‘i] mkhas [pa]

composite-NMLZ TERM join-NMLZ be_good-NMLZ TERM do-NMLZ GEN be_skilled-NMLZ
'[those who are] skilled in good making joining [of words] for composition'

Even if we do not take into account zero nominalization of the last verbal root *mkhas* 'to be skilled', other verbs in this passage are obviously in subordinate syntactic relations of different types with the omission of various grammatical markers. Omission of grammatical markers may be considered acceptable in a poetic text. However, changing the whole formal grammar to ensure correct syntactic analysis of this passage will inevitably cause combinatorial explosions. In this regard, it was decided to model the whole passage as a compound.

According to the created model of Tibetan compounds, reconstructed syntactic relations allow to consider *sdeb-sbyor* and *legs-mdzad* as compound atomic verbal phrases with circumstance (CompoundAtomicVPWithCirc). The CIC CompoundAtomicVPWithCirc was made for a combination of CompoundAtomicVP (verbal phrase within a compound represented by a single verb root morpheme – the head class) and the modifier – CompoundCircumstance, attached on the left. CompoundCircumstance stands for a terminative noun phrase within a compound, consisting of one atom (CompoundAtomicTerminativeNP) and the intersyllabic delimiter (the terminative case marker is omitted as is usual in compounds). The basic class of the nominal component of CompoundAtomicVPWithCirc should be connected by the relation 'to be a relationship object' with the relation 'to have a manner of action or state' – the terminative case meaning.

Thus, for the compound *legs-mdzad*, this relation was established on the basic class of its nominal component *legs-pa* 'being good' – 'any process'. Syntactic relations between *sdeb-sbyor* 'poetry' and *legs-mdzad* 'to do well' are the same as in compound transitive verb phrases (CompoundTransitiveVP), where the first nominal component is a direct object of the second verbal component. In turn, the syntactic relations between *sdeb-sbyor-legs-mdzad* and *mkhas-pa* are the same as those between the components of a noun phrase with genitive compounds (NPGenCompound). These cases of compounds with complex structure are not common, so it was decided not to change immediate constituents of the CICs CompoundTransitiveVP and NPGenCompound, but to create separate classes for compound's groups.

3.6 Annotation error statistics

As mentioned above, the AIIRE corpus manager automatically searches and counts cases of unrecognized fragments, gaps between syntactic structures or their overlaps, and combinatorial explosions. Regular processing of the *Sum-cu-pa* grammar gives us the following statistics of these annotation errors (table 1). The third column also takes into account special cases of described changes, as well as some minor changes, which were not described here.

	1	2	3	4
	Before introducing the ontology	Before the improvements proposed	After the improvements proposed	Current amount
Amount of gaps (tokens)	151	18	7	9
Unrecognized (tokens)	10	0	0	0
Overlaps (tokens)	9	7	0	4
Combinatorial explosions (tokens)	0	0	0	0
Amount of gaps (sentences)	744	323	196	196
Unrecognized (sentences)	26	1	0	0
Overlaps (sentences)	96	59	26	31
Combinatorial explosions (sentences)	7	0	0	1

Tab. 1. The *Sum-cu-pa* grammar processing statistics

The statistics in Tab. 1 takes into account only text processing cases where the syntactic and the semantic analysis were performed simultaneously. Here, we do not provide statistics of annotation errors only for the syntactic mode parsing, because the processing of texts without semantic restrictions at the previous stages of work showed critical ambiguity at the level of syntax.

4 CONCLUSIONS AND FURTHER WORK

The fine-tuning of the automatic morphosyntactic and semantic annotation of the *Sum-cu-pa* grammatical treatise eliminates all unrecognized fragments in the text, almost completely eliminates combinatorial explosions, and significantly reduces gaps in annotation. The remaining breaks are caused by the lack of full semantic annotation of the text (in the syntactic parsing mode the number of gaps is much lower, but the number of versions of parsing becomes unacceptably large).

At the moment, most of linguistic phenomena observed in the text are explained in the current version of formal grammar. We take into account that some of them can be characteristic only for poetic texts or only for texts of a certain period of the Tibetan language development. Further work will mainly include development of semantic annotation of the Sum-cu-pa and completion of work with all the texts of the corpus (of different periods, poetic and prose).

ACKNOWLEDGEMENTS

This work was supported by the Russian Foundation for Basic Research, Grant No. 19-012-00616 Semantic interpreter of texts in the Tibetan language.

References

[1] The Basic Corpus of the Tibetan Classical Language. (2019).

[2] The Corpus of Indigenous Tibetan Grammar Treatises. (2019).

[3] The corpus of Tibetan grammatical works. In Automatic documentation and mathematical linguistics, vol. 49, no. 5, pages 182–191. https://doi.org/10.3103/S0005105515050064.

[4] Wagner, A., and Zeisler, B. (2004). A syntactically annotated corpus of Tibetan. In Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon, pages 1141–1144.

[5] Semantic Roles, Case Relations, and Cross-Clausal Reference in Tibetan. URL: http://www.sfb441.unituebingen.de/b11/ b11corpora.html# clarkTrees.

[6] Grokhovskiy, P., Khokhlova, M., Smirnova, M., and Zakharov V. (2015). Tibetan Linguistic Terminology on the Base of the Tibetan Traditional Grammar Treatises Corpus. In P. Král and V. Matoušek (eds.), Text, Speech, and Dialogue. TSD 2015. Lecture Notes in Computer Science, vol 9302. Springer, Cham.

[7] Dobrov, A., Dobrova, A., Grokhovskiy, P., Soms, N., and Zakharov V. (2016). Morphosyntactic analyser for the Tibetan language: aspects of structural ambiguity. In International Conference on Text, Speech, and Dialogue, pages 215–222. DOI 10.1007/978-3-319-45510-5_25.

[8] Aho, A. V., and Corasick, M. J. (1975). Efficient string matching: An aid to bibliographic search. Communications of the ACM. 18, 6, pages 333–340.

[9] Dobrov, A., Dobrova, A., Smirnova, M., and Soms, N. (2019). Formal Grammatical and Ontological Modeling of Corpus Data on Tibetan Compounds. In Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, vol. 2: KEOD, Vienna, Austria.

[10] Dobrov, A., Dobrova, A., Grokhovskiy, P., Soms, N. (2017). Morphosyntactic Parser and Textual Corpora: Processing Uncommon Phenomena of Tibetan Language. In Proceedings of the International Conference IMS-2017, pages 143–153. DOI 10.1145/3143699.3143719.

[11] Dobrov, A., Dobrova, A., Grokhovskiy, P., Smirnova, M., and Soms, N. (2018). Computer ontology of Tibetan for morphosyntactic disambiguation. In D. A. Alexandrov, A. V. Boukhanovsky, A. V. Chugunov, Y. Kabanov and O. Koltsova (eds.), Digital Transformation and Global Society, pages 336–349, Cham. Springer International Publishing. https://doi.org/10. 1007/978-3-030-02846-6_27.

[12] Dobrov, A., Dobrova, A., Grokhovskiy, P., Smirnova, M., and Soms, N. (2018). Modeling in a computer ontology as a morphosyntactic disambiguation strategy. In P. Sojka, A. Horák, I. Kopecek and K. Pala (eds.), Text, Speech, and Dialogue. TSD 2018. Lecture Notes in Computer Science, vol. 11107, pages 76–83, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-030-00794-2_8.

[13] Grokhovskii, P., and Smirnova M. (2017). Principles of Tibetan compounds processing in lexical database. In Proceedings of the International Conference IMS, pages 135–142. SCITEPRESS. ISBN: 978-1-4503-5437-0. DOI 10.1145/3143699.3143718.

[14] Beyer, S. (1992). The Classical Tibetan Language. State University of New York, New York.

[15] Hill, Nathan W. (2019). Tibetan zero nominalization. Revue d'Etudes Tibétaines, no. 48, Paris.

sciendo
DOI 10.2478/jazcas-2021-0031

ARTIFICIAL HOMONYMY

JAROSLAVA HLAVÁČOVÁ
Institute of the Formal and Applied Linguistics, Charles University, Prague, Czech Republic

HLAVÁČOVÁ, Jaroslava: Artificial homonymy. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 330 – 341.

Abstract: The paper presents a discussion of homonymy of Czech nouns with different or varying genders. The lemmas with this type of homonymy are treated in the new release of the MorfFlex dictionary as separate. We show that the separation of paradigms according to the gender is not only superfluous, but also clumsy, because it forces a choice when making one is not necessary. That is why we call this type of hononymy "artificial".
Keywords: homonymy, polysemy, gender variation, dictionary

1 BASIC CONCEPTS

There are several definitions of homonymy. For the purpose of this paper, we will use the following one:

Homonyms are words with the same spelling but accidentally different meanings.

The definition concerns only one part of homonymy, namely the homography (identical written forms). There is also the homophony (identical pronunciation), but in this paper, only homography will be dealt with under the term homonymy.

There are two terms in this definition that might be a source of misunderstanding.

As for the "words", there are two basic ways how to capture them: as individual wordforms or as lemmas. Artificial homonymy relates to the homonymy of lemmas.

The more difficult term in the definition is the one of "meaning". No unambiguous, simple definition of the "meaning" exists. That is the reason why we will use the term meaning in accordance with "common sense". We consider two lemmas homonymous if their meanings are not connected by any means. In other words, if their spelling is the same only by chance. The example is the lemma *kolej,* which has two independent meanings in Czech: 1. a housing facility for students (college, dormitory), 2. a track or rail.

On the other hand, if a word is used in a figurative meaning, for instance as a metaphor, we consider it "only" polysemous, not homonymous, though we are

aware that the distinction between the two terms is fuzzy.[1] Thus, the lemma *ušák* (1. a hare with big ears, 2. a chair with "ears" resembling a hare, or a pot with big handles resembling ears of a hare) is polysemous, not homonymous, because all the meanings relate to the ears. Another example of the polysemy is using the same proper name for a personal name, as well as for the name of his or her company or firm (e.g. *Albert*). We will discuss individual types of such polysemy later.

If two words with the same spelling belong to different parts of speech, we always consider them homonymous, without regard to their meanings. In other words, in our interpretation, words of different parts of speech have always different meanings.

Such "inter-POS" homonymy is widespread in English, where many words (lemmas) can be used as a verb, a noun, and an adjective. An example is the lemma *house*. The following examples are from the British National Corpus (BNC).[2]
The guard was still in the house (noun).
The practice in medieval times was to house (verb) *all of the grain crops in the barn.*
The house (adjective) *door was locked.*

In the rest of the paper, we will cope with the Czech language only.

1.1 Homonymy in Czech

With its rich morphology, the homonymy is very common in Czech, but not so much in our sense – among the lemmas.[3] Contrary to English, there are only several lemmas that can be used as different parts of speech similarly to the above English example with the *house*. It does not mean that homonymy does not exist. Lemmas with the same spelling and different meanings do exist in Czech. However, we want to show that there is a large set of homonymous lemmas where the homonymy (in the sense presented above) is "artificial". In other words, it is not necessary to call it homonymy as we are convinced that there are not two, but only one word with polysemous property.

There are two basic types of artificial homonymy.

The first one is the homonymy of nouns with different genders, the second one is the homonymy among adverbs, particles, conjunctions, possibly also interjections and prepositions.

The latter type is highly dependent on the definitions of the parts of speech included in the list. This is the reason why we lay aside this type of homonymy. This

[1] The simple distinction between homonymy and polysemy is given by dictionaries – polysemous words usually have a single headword, while homonymous ones are divided into more headwords.

[2] Data cited herein have been extracted from the British National Corpus, distributed by the University of Oxford on behalf of the BNC Consortium. All rights in the texts cited are reserved. The examples are from texts A03, A79 and A0N.

[3] The Czech language is especially rich in the so called morphological homonymy – homonymy among word forms. See the impressive treatise in Petkevič [1].

problem is very complex and its scope extends beyond the possibilities of this paper. Our recent paper deals only with the homonymy of nouns.

2 MOTIVATION – NOUNS WITH VARYING GENDER

The basis of our study is the Czech morphological dictionary MorfFlex CZ ([2], [3]), examples were taken from the corpora of the SYN[4] [4] and Aranea[5] [5] series. For its latest edition, version 2.0, several principles were applied to make the content of MorfFLex consistent (see [2]). One of those principles is the "Principle of unique paradigm" saying that there are no two identical paradigms (sets of lemma-tag pairs) in the dictionary. It means that every paradigm has only one lemma, even if it has more meanings. This rule was adopted for the reason of simplicity. The lemma *kolej* presented above, is a typical example.

Another basic principle – "Principle of morphological differentiation" – implies that nouns with different genders are different.

The Czech language has three genders – masculine, feminine, and neuter. The masculine gender may be animate or inanimate. These two subgenders have partially different inflections. It was probably the reason why they are usually considered two separate genders in the field of NLP. Thus, in most of the Czech morphological tagsets, there are codes for 4 genders: masculine animate, masculine inanimate, feminine, and neuter.

The great majority of Czech nouns have a single gender within their paradigm. However, there are nouns with varying genders.

There are two basic ways how to describe that situation morphologically. The way adopted by the authors of the new version of the MorfFlex (see above) was the division of the paradigm with the varying gender into more paradigms, each having all the wordforms of a single gender. In such way, the paradigms became separated, each represented by its own lemma. As both lemmas have the same spelling, they become homonymous. Technically, in the morphological dictionary, they are distinguished by means of a numerical index added to the lemma. See the example of the word *kredenc* in Tab. 1.

In our view, this solution is superfluous and the resulting homonyms are artificial. We suggest another solution – rejection of the part of the Principle of morphological differentiation concerning the noun gender. There is no reason why wordforms within one paradigm should have only one gender. It is even in contradiction with the reality.

Let us illustrate both approaches on an example with the lemma *kredenc*. In the present version of the dictionary, we have *kredenc-1* with the masculine inanimate

[4] Accessible at: http://www.korpus.cz.

[5] Accessible at: http://unesco.uniba.sk/aranea/.

gender and *kredenc-2* with the feminine gender. The meaning of the both is the same – a cupboard.

If we admitted both genders in the same paradigm, there would be no need to have two lemmas. The set of wordform-tag pairs will be the union of the pairs from both paradigms in the former approach (see Tab. 1).

There are several more types of lemmas divided under that principle. They will be discussed in the following sections.

In the field of NLP, the only thing that should not be violated is the Golden rule of Morphology ([2], [6], [7]) saying that every combination of a lemma and a morphological tag must not be represented by more than one wordform. If the two wordforms with the same lemma differ in their gender, even if the rest of the morphological features is identical, their tags do differ, which is a sufficient condition for meeting the Golden rule requirement. However, there are some issues that have to be mentioned and resolved.

2.1 Lemma of a varying gender paradigm

If we merge the paradigms of artificial homonyms with a varying gender, a question may arise what will the gender of its lemma be? The spelling of the lemma is unique, but within a merged paradigm, it can be assigned two genders, depending (only) on the context.

	kredenc				
kredenc-1	*kredence*	NNIP1-----A----		*kredence*	NNFP1-----A----	*kredenc-2*
	kredenců	NNIP2-----A----		*kredencí*	NNFP2-----A----	
	kredencům	NNIP3-----A----		*kredencím*	NNFP3-----A----	
	kredencum	NNIP3-----A---6				
	kredence	NNIP4-----A----		*kredence*	NNFP4-----A----	
	kredence	NNIP5-----A----		*kredence*	NNFP5-----A----	
	kredencích	NNIP6-----A----		*kredencích*	NNFP6-----A----	
	kredenci	NNIP7-----A----		*kredencemi*	NNFP7-----A----	
	kredencema	NNIP7-----A---6		*kredencema*	NNFP7-----A---6	
	kredenc	NNIS1-----A----		*kredenc*	NNFS1-----A----	
	kredence	NNIS2-----A----		*kredence*	NNFS2-----A----	
	kredenci	NNIS3-----A----		*kredenci*	NNFS3-----A----	
	kredenc	NNIS4-----A----		*kredenc*	NNFS4-----A----	
	kredenci	NNIS5-----A----		*kredenci*	NNFS5-----A----	
	kredenci	NNIS6-----A----		*kredenci*	NNFS6-----A----	
	kredencem	NNIS7-----A----		*kredencí*	NNFS7-----A----	

Tab. 1. Merged paradigms of the lemma *kredenc-1* and *kredenc-2*

In the sentence (1), the adjective wordform *malou* is described by the lemma *malý* (small) and the tag AAFS4----1A---- (the 3rd position F says that it is feminine), while in the sentence (2) the lemma is the same, *malý*, but the tag differs in the code I for the gender (masculine inanimate) at the position 3: AAIS4----1A----. From the forms of the adjective, the gender of the noun *kredenc* is deduced. The noun *kredenc* can have the same lemma in both examples, but its tags will differ in gender. Both sentences mean the same: 'We have a small cupboard.'

(1) *Máme malou kredenc.* (feminine)

(2) *Máme malý kredenc.* (masculine inanimate)

The answer is simple. There is no need to assign any gender to the lemma. The morphological tag is not part of the lemma. The lemma is a wordform in nominative (usually singular, but there are also pluralia tantum – see later). Its written form can be described with two tags, which differ in gender, but the lemma itself is unique.

2.2 Gender of undistinguishable wordforms

Another problem could be assigning a gender to a wordform with an undistinguishable gender in the given context. An example is the sentence (3), where it is not clear which gender is the right one:

(3) *Máme kredenc.* ('We have a cupboard.')

It differs from the previous examples (1) and (2) by the missing adjective – there is no clue how to decide about the gender of the wordform *kredenc*. Thus, in the present setting of the dictionary, with two lemmas for *kredenc*, it is necessary to choose one of them arbitrarily. If we reject the artificial homonymy, the lemma assignment is easy. However, the necessity of a choice will not disappear. We still have to choose between the two genders, more precisely – between the tags with different genders, because the lemma is now unique – *kredenc*.

In fact, the necessity of choice is the same. Evidently it is not important, the gender can be assigned randomly in such cases. Sometimes, an objection appears, that it is necessary that the gender should be the same throughout a single text. With sofisticated advanced automatic tools this is achievable, or will probably be soon. However, very often even human authors are not consistent throughout a single text. That is why automatic tools need not be consistent, either. It follows that the selection of an appropriate gender for such an occurrence can really be arbitrary.

There are several ways how to decide on a unique solution for each word in a context without any clue for its gender. The gender selection might be random, or according to a criterion. The simplest solution would be an ordering according to a gender preference, the same for all the lemmas. The most natural would probably be this one: M, I, F, N. The rule for the gender assignment would be: If the gender of a wordform cannot be decided from the context, pick one which is the leftmost in the above list. According to that simple rule, the gender of *kredenc* from the example (3) would be I (masculine inanimate).

3 GENDER COMBINATIONS AMONG NOUNS

Let us have a look at possible combinations of genders in the set of all artificial homonyms from the morphological dictionary MorfFLex CZ 2.0.

The following sections will have the names according to codes of genders: M masculine animate, I masculine inanimate, F feminine, N neuter.

3.1 MN

There is 15 lemmas of that kind in MorfFlex, two of them being problematic, possibly wrong. The majority of them (9) are lemmas ending wigh *–e* or *–ě*. They are old words denoting mainly members of nobility (*hrabě* – 'earl', *markrabě* – 'margrave' etc.). They have very unusual morphology for masculine gender in Czech. All these words have the paradigm typical for the neuter gender. However, according to contexts, both genders are plausible. In corpora, we can find expressions or sentences such as:

hrabě hohenembský (masculine) 'Earl of Hohenemb';

V 18. století ji vlastnili *hrabata* z Bubnu (masculine) 'In the 18th century, it was owned by the earls of Buben';

S nimi spříznění *hrabata* Stadničtí ho drželi až do počátku druhé světové války. (masculine) 'earls of Stadnicty, related to them, owned it until the outbreak of World War II';

V návštěvních knihách pak čteme další jména, svědčící o tom, že zámek navštěvovala knížata, *hrabata* (neuter) 'In guest books, we read further names, which shows that the castle was visited by counts and earls';

Hrabata Desfours-Walderodové na Dřínově byla podle vypravování typická *hrabata*, jak je známe z anekdot a divadelních frašek. (neuter) 'The earls of Desfours-Walderod on Dřínov were reputed to be typical earls, as we know them from anecdotes and theatrical farces'.

There are two more words with a typical neuter ending: *pako* ('nitwit') and *libero* ('libero'). The former one is colloquial, appearing often in a nonstandard context, the latter one belongs to sport slang. Their gender really varies but we can see the tendency to use neuter gender in singular and masculine gender in plural (Tobě *podobní paka* mě opravdu nepřekvapí. (masculine animate) 'Nitwits like you won't really surprise me.')

Finally, there is the word *cockney*. In the neuter gender, it means a dialect, in the masculine gender, a man speaking in that dialect. This particular gender selection may be subject to discussion but in any case, *cockney* is one word and as such should be represented with a single lemma. The same is true for the whole group.

3.2 IN

All the words in this group are loanwords. Some of them originally come from ancient Latin or Greek (*ostrakon* – 'ostracon' – a piece of ancient pottery) where it has the neuter gender, but after becoming part of the Czech vocabulary, people

started to decline them according to their formal ending, which resembles the Czech masculine gender. Thus, in texts, both declensions appear.

Some of the words are loanwords with unstable declension (*blues*, *interview*).

There is only one word in this group where the separation is reasonable. *House-1* is the Czech word meaning a gosling (young goose) and *house-2* is a loanword with the English meaning house, but with the Czech declension. Here, strict separation of genders is justified, as the identical spelling is only incidental. They are true homonyms.

3.3 IM

The combination of animate and inanimate masculine declension contains 571 nouns. There are four main semantic groups:

A: inanimate lemmas used in figurative meaning for masculine animate persons. For instance *truhlík* (‘box’), *věchýtek* (‘bundle of straw’), *hajzl* (‘toilet’ – vulgar), *klenot* (‘jewel’);

B: animate lemmas used in figurative meaning for inanimate things: špaček (‘starling’), *hlemýžď* (‘snail’), *ušák* (‘rabbit’ or ‘hare’), žralok (‘shark’);

C: a tool or a person doing the same thing as the tool: *kompilátor* (‘compilator’), *konstruktor* (‘constructor’), *komunik*átor (‘comunicator’), *dělič* (‘divider’), *držák* (‘holder’);

D: other nouns used in both genders with the same or very similar meaning: *tenor* (‘tenor’ as a man or as a voice), *exot* (‘freak’), člen (‘member’), *solitér* (something or someone appearing uniquely).

The groups A and B contain lemmas that appear often in an expressive metaphorical meaning. It is the reason why we cannot proclaim their meaning independent. They are polysemous, but not homonymous.

The group C contains nouns denoting either men, or tools/means of an activity (*constructor* is someone or something that constructs, *držák* (‘holder’) is someone or something that holds, etc.). There are contexts in which it is even not possible to guess the correct gender. In such case, however, the gender should probably be consistent throughout a single text. Thus, the precedence rule would not be appropriate here.

The group D contains words with very similar meanings, the gender of which often cannot be distinguished even in some contexts. For instance in the sentence:
Tenor se většinou ... vůbec neprosadí. (‘The tenor usually doesn’t succeed at all.’)
it is not clear, if the *tenor* is a singer (animate), or his voice (inanimate).

There is one word in the group IM that is really homonymous: *rys*. Its two meanings are not connected (animate 1. lynx, inanimate 2. feature), the same spelling is accidental. In this case, the two paradigms with two lemmas are reasonable. In the inanimate gender, the word *rys* has more meanings, but according to the Principle of morphological differentitaion, there is only one lemma with that gender, regardless of more meanings.

3.4 FI

Nouns having both genders, feminine and masculine inanimate, are typical words with varying gender. One of them – *kredenc* – has been already discussed. Many Czech names of geographical objects, villages and towns, belong to this group. Very often only inhabitants of the place know the correct gender, as it is usually a matter of dialect or tradition. A famous example is the Moravian town of *Olomouc,* but there are many other (*Bubeneč, Černíč*, *Dobrovíz, Radom*). There are even different villages with the same name but different genders, according to the local tradition, but it is not reasonable to have two lemmas for them, as the usage varies in those cases, too.

This type of gender variation comprises also nouns that appear only in plural – the so called pluralia tantum. It is usually impossible to guess their gender from the lemma. There are only several grammatical cases, from which it is possible to deduce their gender. However, they are very often not unique. In other words, their gender varies, too.

Compare the paradigms of the lemma *varhany* ('organ' – musical instrument) in Tab. 2. In the two left columns, there are wordforms and their cases. The last two columns contain their frequency in the corpus Araneum Bohemicum IV Maximum of the Aranea series. The black lines mark wordforms that are identical for both genders in the given case. White lines contain only feminine wordforms, the grey lines contain only masculine gender.

Wordform	case	gender F	gender I
varhany	1 (nom)	16 874	1 276
varhan	2 (gen)	10 089	3 360
varhanů	3 (dat)	---	62
varhanám	3 (dat)	556	---
varhanům	3 (dat)	---	170
varhany	4 (acc)	9 988	1 170
varhanech	6 (loc)	---	108
varhanách	6 (loc) non-stand.	1 000	31
varhanami	7 (instr)	1 567	---
varhanama	7 (instr) non-stand.	5	28
varhany	7 (instr)	---	416

Tab. 2. Gender distribution across all wordforms of *varhany* according to the annotation in the corpus Araneum Bohemicum IV Maximum

The black lines are unfathomable. There is no reason why to assign different gender to those wordforms, as there cannot be a single clue for their distinction in

any context. On the other hand, that division causes no problem. It is only strange and cannot be explained.

The more natural solution could be assigning a single gender to the whole paradigm. Where there are different forms for a particular case (dative, locative, instrumental), the wordforms could be considered as variants. Nevertheless, preserving the current state with different genders is also reasonable. The only change should be merging all the wordforms under a single lemma of *varhany*. Having two lemmas, *varhany-1* and *varhany-2,* does not make any sense. The same applies to all pluralia tantum from this group, including, again, the proper names (*Lažánky, Sudety*).

Generally, in the case of pluralia tantum with the gender varying between F and I, the gender is not important at all. It has no influence on any type of agreement. That is why it is not necessary to assign two genders to them. It is reasonable to choose one according to etymology, dialect, or any other clue, or even randomly, and to proclaim the forms resembling the other gender inflectional variants.

Another solution would be selection of a single gender for those wordforms that do not differ. Wordforms with a "visible" gender can keep the different gender. In any case, it is not necessary to create a different lemma for them.

3.5 FM

There are 206 lemmas having this combination of genders. As the masculine animate gender is involved, it is clear that feminine "homonyms" will also denominate living creatures, persons, or animals. There are 78 lemmas ending with –í type (*strojvedoucí* – 'train driver'). These nouns follow the soft adjective declension where the gender manifests itself only in several combinations of case and number.

Another large group (121) are nouns ending with *–a*. These nouns are usually semantically related as they are sort of expressive nicknames for persons, both men and women: *pápěrka* ('weakling'), *sirota* ('orphan'), *trouba* ('simpleton', but also 'an oven'), etc. Many of them are derived from general words and their usage for a person denomination is a sort of a metaphor, similar to the group A in the section IM.

The rest are loanwords with indefinite gender: *hippie*, *pair*, (*super*)*star*, *sfinx*, *okapi*.

There are two more old words, namely *choť* ('spouse') and *sršeň* ('hornet'). The former one can refer to both a man or a woman, while the latter one appears ambivalently – as a scientific name it is feminine, but it is commonly used as masculine.

3.6 FN

This group contains mainly loanwords. Examples: *panorama, scifi, promile*. However, the most interesting (and problematic) words are three old Czech words,

namely *oko* (‘eye’), *ucho* (‘ear’), and *dítě* (‘child’) and their derivatives (for instance *biodítě* ‘bio-child’). They have an unusual morphology, because their varying gender has a system; they are neuter in singular and feminine in plural.

The solution adopted in the new MorfFlex is the division of the paradigm according to the grammatical gender, which corresponds with the grammatical number. Thus, we have the lemma *dítě-1* having only neuter wordforms in singular, and lemma *dítě-2* with only feminine wordforms in plural.

Oko and *ucho* are even more complicated. They have the regular declension in neuter for plural, too, but only for the figurative meanings. When speaking about an organ of a vision or hearing, the gender changes to feminine in plural. In this case, the results are lemmas *oko-1, ucho-1* with the regular declension and the both numbers, and *oko-2, ucho-2* that have only plural feminine wordforms in their paradigms.

All those irregular words could be captured simply within a single paradigm with a single lemma and a varying gender.

3.7 FIN

There is one word that appears in texts in three genders. It is the loanword *image* that became quite popular but as its ending, and the disagreement of its written and pronounced form, does not correspond with any Czech pattern, people use it in all three genders.

3.8 FMN

There are two words in this group: *budižkničemu* (‘good-for-nothing’) and *rukojmí* (‘hostage’). Their belonging to more genders follows from the fact that their endings are not typical for any nominal gender. The gender in a particular context can be derived from the agreement rules. Whenever it is not possible, the gender M seems to be the most appropriate. In any case, all the paradigms can (should) be merged into a single one, as the meanings are the same for all the genders.

3.9 Gender combinations among foreign proper nouns

Foreign proper names appear very often in the language data. If they denote persons, they can get the gender according to the sex of the person – either masculine animate or feminine. If he or she has a company with the same name, we usually assign it the masculine inanimate or neuter gender. Sometimes, names of persons are the same as geographical names, with different genders. Thus, the same proper name can have all genders. In such cases, the gender is often subspecified with the code X saying that any gender is possible. For a particular noun, every possible combination of genders can appear in corpora, including the subspecified X.

This situation is visible in MorfFlex CZ 2.0. In fact, the genders were assigned to foreign proper names according to findings in the data, namely the new PDT-C

corpus [3], that was manually annotated. On the one hand, it is nice that the dictionary is in agreement with the corpus, on the other hand, this solution is not general. We are convinced that a generalization should be made so that foreign names may fit in any context in which they appear in the future. Thus, a single lemma and a maximally subspecified morphological tag would be the most general solution. For names that can undergo Czech inflection, the paradigm may include all the appropriate wordforms with their tags, no matter which gender the wordforms will acquire.

Moreover, adding foreign proper names to the lexicon is a neverending task. Guessers should be used instead of increasing the dictionary with this type of words.

4 CONCLUSION

We discussed polysemous nouns and their treatment in the morphological dictionary MorfFlex CZ 2.0. If they have the same gender, they are now treated as a single lemma with a unique paradigm. Where the genders differ or, they are treated as homonyms. It follows that there are two (or even more) paradigms, each of them with a unique gender, represented by several lemmas distinguished by means of numerical indexes. We call these homonyms artificial, because we are convinced that they are not homonyms at all. Meanings of the great majority of them are interconnected, if not even the same (esp. for lemmas with varying gender).

We presented an overview of possible gender combinations, using the dictionary MorfFlex and Czech corpora, to show that dividing such lemmas according to their gender is not necessary. We suggest all the wordforms of possibly more meanings merge into a single paradigm with a unique lemma. It will make morphological annotation simpler. Also, maintenance of the lexicon will become easier, especially with respect to foreign words, as there will be no need to add and to number new lemmas if they appear with a different gender of a foreign word in future data.

ACKNOWLEDGEMENTS

This contribution was supported by the project of the Ministry of Education, Youth and Sports of the Czech republic: LINDAT/CLARIAH-CZ (LM2018101): Digital Research Infrastructure for the Language Technologies, Arts and Humanities, 2019–2022.

References

[1] Petkevič, V. (2016). Morfologická homonymie v současné češtině. NLN, Praha, 588 p.

[2] Mikulová, M., Hajič, J., Hana J., Hanová, H., Hlaváčová, J., Jeřábek, E., Štěpánková, B., Vidová Hladká, B., and Zeman, D. (2020). Manual for Morphological Annotation, Revision for the Prague Dependency Treebank – Consolidated 2020 release (technical report).

[3] Hajič, J., Hlaváčová, J., Mikulová, M. et al. (2020). MorfFlex CZ 2.0, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3186.
[4] Hnátková, M., Křen, M., Procházka, P., and Skoumalová, H. (2014). The SYN-series corpora of written Czech. In Proceedings of the Ninth International Conference on Language Resources and Evaluation, pages160–164. Reykjavik: ELRA.
[5] Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In P. Sojka, A. Horák, I. Kopeček and K. Pala (eds.), Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, Proceedings. LNCS 8655. Springer International Publishing Switzerland, pages 257–264.
[6] Hlaváčová, J. (2017). Golden rule of morphology and variants of wordforms. Jazykovedný časopis, Bratislava, Slovak Academic Press. DOI 10.1515/jazcas-2017-0024.
[7] Hlaváčová, J. (2009). Formalizace systému české morfologie s ohledem na automatické zpracování českých textů. Ph.D. thesis, FF UK, Praha.

sciendo
DOI 10.2478/jazcas-2021-0032

TYPOLOGICAL PROFILING OF ENGLISH, SPANISH, GERMAN AND SLOVAK: A CORPUS-BASED APPROACH

JAKOB HORSCH
Catholic University of Eichstätt-Ingolstadt, Eichstätt, Germany

HORSCH, Jakob: Typological profiling of English, Spanish, German and Slovak: A corpus-based approach. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 342 – 352.

Abstract: Inspired by earlier work on typological profiling of English by Benedikt Szmrecsányi and Bernd Kortmann ([1], [2], [3]), this paper investigates the typological profiles of English, Spanish, German, and Slovak, applying Szmrecsányi and Kortmann's methodology of calculating a SYNTHETICITY INDEX and an ANALYTICITY INDEX based on 1,000-word corpus samples. The results show that Szmrecsányi and Kortmann's methodology is replicable, and confirm claims in the literature about degrees of analyticity and syntheticity of these languages. Instead of a simple analytic-synthetic continuum, Szmrecsányi and Kortmann's "typological space" [3] is used to visualize results, showing that languages can be both synthetic *and* analytic to varying degrees.

Keywords: typological profiling, syntheticity index, analyticity index, typological space, English, German, Spanish, Slovak, corpus samples

1 INTRODUCTION

In morphological typology, the terms *synthetic* and *analytic* are widely used to describe languages based on their morphosyntactic properties. Accordingly, languages are characterized "as rather analytic […] or as rather synthetic" [1]. English, for example, has been referred to as "analytic to a very high degree" [4], and Slovak is considered a synthetic, or inflecting language: It is described as "*výrazne flektívna*" ('significantly inflecting') [5] or "*Slovenčina je prevažne flektívny jazyk*" ('predominantly inflecting') [6]. In-between the analytic and synthetic extremes are languages like Spanish, which has "retained a large number of synthetic verb forms while undergoing a radical change towards analyticity in the domain of nouns and adjectives", and German, where "the verb phrase […] is one of structural and lexical analyticity […] combined with a fairly high degree of syntheticity in the maximally governing finite verb" [4].

But just *how* synthetic or analytic are languages? Statements like 'analytic to a very high degree', 'predominantly inflecting', 'radical change towards analyticity', and 'fairly high degree of syntheticity' remain somewhat vague. Addressing this issue, the main objective of the present study is to determine and compare the typological (i.e., morphosyntactic) profiles of four Indo-European

languages from across the spectrum that ranges from analytic to synthetic: English, Spanish, German, and Slovak. Following the methodology of Szmrecsányi and Kortmann [3], degrees of analyticity and syntheticity will be calculated based on random samples drawn from corpora.

Before proceeding, a brief[1] overview of the history of the terms *synthetic* and *analytic* will be provided – after all, they "have a long and venerable tradition in linguistics" [1]. They were coined in the early 19th century by August Wilhelm von Schlegel[2] [7], whose "simple binary classification" [8] was refined by Sapir [9] into a "scalar concept of syntheticity" [8]. Building on Sapir's work, Greenberg devised a mathematical formula for calculating a "synthetic index", which he defined as "the ratio M/W where M equals morpheme and W equals word", such that "[a]nalytic languages will give low results on this index, synthetic higher" [10].

Although Greenberg's formula is very appealing because of its simplicity, reality is more complicated, which is reflected in the fact that the terms 'analytic(ity)' and 'synthetic(ity)' are not used consistently in the literature. Szmrecsányi, for example, points out "terminological confusion" [1], and Schwegler laments a "vagueness of terms" [8] in this context. Therefore, precise definitions are in order at this point.

The approach to analyticity and syntheticity adopted in the present study is that of Szmrecsányi and Kortmann, who have done some groundbreaking work. Inspired by Greenberg's syntheticity index, Szmrecsányi and Kortmann defined "overt *grammatical syntheticity*" as "the text frequency of bound grammatical markers" [3] (emphasis in the source), and "overt *grammatical analyticity*" as "the text frequency of free grammatical markers" [3] (emphasis in the source), adding a dimension that had largely been ignored previously.[3] Thus, languages could be profiled in more detail by providing measurements of syntheticity *and* analyticity. Whereas Greenberg's synthetic index was used to describe degrees of syntheticity *versus* analyticity, with Szmrecsányi and Kortmann's approach it was now possible to describe degrees of syntheticity *and* analyticity.

As the terms *grammatical* syntheticity and *grammatical* analyticity indicate, Szmrecsányi and Kortmann [3] focus on the marking of grammatical information, disregarding lexical processes such as derivation and compounding. Furthermore, Szmrecsányi and Kortmann's approach is only concerned with *overt* grammatical marking, that is, the *presence* of morphemes, ignoring phenomena such as null

[1] The reader is referred to chapter 1 in Schwegler [8] for a detailed overview (as pointed out by Szmrecsányi [1]).

[2] Schlegel originally used the terms 'synthetic' and 'analytic' to distinguish between the evolutionary stages of inflectional languages, as noted by Askedal [4].

[3] Although, as Szmrecsányi notes [1], the need for an analyticity index was noted as early as the 1980s by Kasevič and Jachontov [11].

marking, or zero morphemes.[4] They do, however, take into account "allomorphs including ablaut phenomena […] and other nonregular, yet clearly bound grammatical markers" such as suppletion [1]. Szmrecsányi and Kortmann's precise definition[5] of grammatical analyticity and grammatical syntheticity is thus as follows:

> "[*F*]*ormal grammatical analyticity* [is defined] as comprising all those coding strategies where grammatical information is conveyed by free grammatical markers, which we in turn define as function words that have no independent lexical meaning. Conversely, we take *formal grammatical syntheticity* to comprise all those coding strategies where grammatical information is signaled by bound grammatical markers." [3] (emphasis in the source)

Based on this definition, Szmrecsányi and Kortmann devised the following two formulas, which will be applied in the present investigation:

> (1) "The *analyticity index*: the ratio of the number of free grammatical markers in a sample (F) to the total number of words in the sample (W), normalized to a sample size of 1,000 tokens. Hence: ANALYTICITY INDEX = f/w × 1,000." [3] (emphasis in the source)

> (2) "The *syntheticity index*: the ratio of the number of words in a sample that bear a bound grammatical marker (B) to the total number of words in the sample (W), normalized to a sample size of 1,000 tokens. Hence: SYNTHETICITY INDEX = b/w × 1,000." [3] (emphasis in the source)

Szmrecsányi and Kortmann's work focused on English and its varieties, e.g., intra-lingual variation in English across different registers [1], comparing Learner Englishes to L2 varieties of English [3], and tracing the diachronic evolution of English [2]. However, what is of much greater interest in the context of cross-linguistic typology is what appears to have been more of a by-product of one of their studies. To "investigate the issue of substrate effects" on Learner Englishes [3], Szmrecsányi and Kortmann included the following six European languages: Bulgarian, Czech, French, German, Italian, and Russian. Table 1 provides an overview of their results: the analyticity and syntheticity indices of these languages, based on 1,000-word corpus samples.

The numbers correspond to what most linguists would intuitively predict. Low syntheticity index (SI) scores were determined for English (SI: 197) and French (SI:

[4] Incidentally, most grammars that were consulted in the context of the present study explicitly reject the concept of a zero morpheme (e.g. the approaches of the Real Academia Española's [12] and the *Duden*, a grammar of German [13]). However, some do not (e.g. Dvonč et al. [5] and Oravec [6]).

[5] Note that Szmrecsányi and Kortmann's definitions are "strictly *formal* […] and not semantic" [1] in nature. Thus, the multiple meanings of portmanteau morphs are disregarded in calculating the syntheticity index.

153), which are analytic languages (as noted by, e.g., Oravec [6]). Conversely, Czech (SI: 683) and Russian (SI: 670), two synthetic languages, have the highest SI scores and the lowest analyticity index (AI) scores (Czech AI: 334; Russian AI: 300). English (AI: 427) and French (AI: 439) score high in this regard, and German (SI: 301, AI: 436), Bulgarian (SI: 394, AI: 372) and Italian (SI: 250, AI: 458) cover the middle ground.

Language	Analyticity index	Syntheticity index
(British) English[6]	427	197
Bulgarian	372	394
Czech	334	683
French	439	153
German	436	301
Italian	458	250
Russian	300	670

Tab. 1. Analyticity and syntheticity index values of various European languages, data from Szmrecsányi and Kortmann [3]

Note, however, that these results are based on "comparatively small corpora" consisting of "approx[imately] 10,000 words of running text each […] sampling newspaper prose" [3]. Furthermore, Szmrecsányi and Kortmann had part-of-speech (POS) annotation carried out by a different person for each language, "typically by native speakers" [3]. While there are good reasons for recruiting native speakers, who can be expected to have sound knowledge of their L1 languages, Szmrecsányi and Kortmann's approach comes with the disadvantage that there might have been inconsistencies in coding, affecting the comparability of their results.

The present study, apart from shifting Szmrecsányi and Kortmann's focus on synchronic and diachronic *intra*-linguistic variation in English to *inter*-linguistic variation, seeks to address these issues by using random samples from much larger corpora and having data annotation carried out by the same researcher. The languages chosen were English, Spanish, German, and Slovak, based on the following considerations: (1) To test the feasibility of applying Szmrecsányi and Kortmann's methodology to further languages (i.e., Spanish and Slovak); (2) to see whether Szmrecsányi and Kortmann's study could be replicated with regard to English and German; (3) because English and Slovak fall on the opposite ends of the analytic-synthetic continuum, with Spanish and German covering the middle ground; and (4) because the author is fluent in all four of these languages, so they could be coded with a high degree of consistency.

[6] Szmrecsányi and Kortmann actually investigated three different registers of the British National Corpus (BNC), including conversation, university essays, and school essays. The value provided here is that of university essays [3], as this register appears to be the most comparable to the newspaper prose that was sampled for the other European languages.

The main **objective of the present study** is, thus, to **calculate, following Szmrecsányi and Kortmann's methodology, the analyticity and syntheticity indices of English, Spanish, German, and Slovak**. The data that will serve as a basis for these calculations consists of 1,000-word random samples extracted from four *Sketch Engine* corpora [14], [15]. Since these corpora were compiled by the same researchers, there should be a high degree of comparability between languages.

The **expected results** are as follows: Based on Szmrecsányi and Kortmann's findings, **English should score high in terms of analyticity and low in syntheticity**. **Slovak**, a synthetic language, should have **scores similar to its close relative Czech**[7], for which Szmrecsányi and Kortmann determined a **low analyticity index and a high syntheticity index** (see table 1). The analyticity and syntheticity indices of **German**, for which Szmrecsányi and Kortmann determined an AI of 436 and an SI of 301, as well as **Spanish**, for which no such indices have been calculated so far, should fall **somewhere in-between English and Slovak**.

2 DATA AND METHODOLOGY

The main objective of this study, as outlined in section 1, is to determine the analyticity and syntheticity indices of English, Spanish, German, and Slovak. The calculations are based on data samples from the Slovak Web 2011 (skTenTen11), the Spanish Web 2018 (esTenTen18), the German Web 2013 (deTenTen13), and the English Web 2015 (enTenTen15) corpora, which were queried using Sketch Engine's online interface[8] in July and December 2020; these corpora were chosen for reasons of comparability.[9] CQL queries were used to extract all words from each corpus, excluding punctuation and other symbols.[10] Subsequently, a 1,000-word sample was drawn by means of Sketch Engine's "get a random sample" function. These samples were downloaded as CSV files and annotated in *LibreOffice Calc* [17], and then saved as ODS files.[11] Each token from the random sample was subsequently annotated for the following variables:

- FUNCTION WORD (levels: TRUE, FALSE)
- NUMBER OF BOUND GRAMMATICAL MARKERS (levels: 0, 1, 2...)

[7] Actually, the two languages are so closely related to each other that they are generally considered to form a "dialect continuum" [16].

[8] Access to Sketch Engine (https://app.Sketch Engine.eu) was generously made available to the Catholic University of Eichstätt-Ingolstadt through the ELEXIS Program (https://elexis.is).

[9] As one reviewer noted, the Slovak National Corpus could have been used to achieve a higher degree of representativeness. However, in the interest of obtaining comparable samples, it was decided to use skTenTen11, which is more similar to the other corpora in that it contains texts from the web only and was compiled by the same researchers.

[10] The full CQL expressions are contained in the files at the research project's OSF repository (see below).

[11] All files – including the original CSVs downloaded from Sketch Engine and the manually coded ODS files – are available at the following OSF repository: https://osf.io/9w3u5/.

Annotation and morphological segmentation to determine the number of bound grammatical markers was carried out based on the following standard grammars:

- English: *A Comprehensive Grammar of the English Language* [18],
- German: *Duden: Die Grammatik* [13],
- Spanish: *Nueva gramática de la lengua española* [12],
- Slovak: *Morfológia slovenského jazyka* [5]; *Morfológia spisovnej slovenčiny* [6].

The coding of the variable FUNCTION WORD, which involved checking whether the word token was "synsemantic", i.e., with "no independent lexical meaning" [1], was greatly facilitated by POS annotation in the corpora. Nevertheless, each token was manually checked, as corpora have been known to contain erroneous tags. Each word token with a POS tag corresponding to closed word classes loaded on to the analyticity index of the respective language. This included prepositions, pronouns[12], determiners, conjunctions, modal/auxiliary verbs, negators, primary verbs in auxiliary function (English), the infinitive marker *to* (English), and particles.

Next, the number of bound grammatical markers (i.e., affixes) was counted. To do so, each word token was segmented using the paradigmatic substitution test to determine whether a morpheme carried meaning. Thus, the Slovak word *pripravila* 'she prepared' would be segmented into three morphemes; a stem (*pripravi-*) and two grammatical affixes that indicate past tense (*-l-*) and gender (*-a*). Regarding English, this was a rather simple undertaking, since no more than one grammatical suffix can attach to a word at a time. In the case of Spanish, German, and Slovak, however, matters were more complex:

- In German, there are certain noun inflection classes whose dative plural endings can be segmented into two separate morphemes, e.g., *den Tag-**e***$_{\text{PL}}$*-**n***$_{\text{DAT}}$ [19]. Also, certain past tense forms of verbs have two segmentable affixes, e.g., *such-**te***$_{\text{PST}}$*-**st***$_{\text{2SG}}$ [19]. Circumfixes, e.g., ***ge***$_{\text{PST}}$*-sag-**t***$_{\text{PST}}$, however, were only counted as one morpheme, as were combinations of *umlaut* with suffixation, e.g., *der Turm*$_{\text{SG}}$ → *die T**ü**rm-**e***$_{\text{PL}}$;
- In Romance languages including Spanish, "number and gender marking on nouns and adjectives is […] typically suffixal" [20], so that these word classes can carry up to two distinct suffixes, e.g., *niñ-**o***$_{\text{M}}$*-**s***$_{\text{PL}}$.[13] Similarly, inflected Spanish verbs can carry up to two suffixes [12], e.g. *cantá-**ba***$_{\text{IMP}}$*-**mos***$_{\text{1PL}}$;[14]

[12] Although note that some Slovak grammars exclude pronouns from the group of synsemantic words [6].

[13] Concerning the approach to the grammatical gender morpheme in Spanish in the present paper, cf. [12].

[14] Note that according to the *Nueva gramática de la lengua española*, inflected verbs in Spanish can actually be segmented into up to four components: root; thematic vowel; tense, aspect and mood marker; person and number marker [12]. In the present study, however, the thematic vowel was disregarded, as it does not carry any meaning [12], making its morpheme status (which in the present study is defined as 'the smallest *meaning-bearing* unit') disputable.

- Slovak superlative adjective forms can be segmented into three affixes (comparative, superlative, and gender), e.g. ***naj***$_{SUP}$-*siln*-***ejš***$_{COMP}$-***ia***$_{F}$.

Thus, in Spanish, German, and Slovak, one word token could load on to the syntheticity index multiple times. Once the data was annotated, the "two Greenberg-inspired index values" [3] were calculated according to Szmrecsányi and Kortmann's formulas (see section 1).

3 RESULTS AND DISCUSSION

After annotating the 1,000-token samples, the analyticity and syntheticity indices were determined for each language. Table 2 presents the results:

Language	Analyticity index	Syntheticity index
English	395	210
Spanish	423	410
German	458	517
Slovak	355	595

Tab. 2. Analyticity and syntheticity indices of English, Spanish, German, and Slovak

A first glance at table 2 confirms the expectations outlined in section 1. English scores low in syntheticity, with an SI of 210 – that is, out of 1,000 words, 210 bear a grammatical marker. In contrast, Slovak has the highest syntheticity score: the 1,000-word sample contained 595 inflectional morphemes. Spanish (SI: 410) and German (SI: 517) are in-between these two extremes. Slovak has the lowest AI score (AI: 355), i.e., only 355 function words were found in the 1,000-word sample. Interestingly, Spanish (AI: 423) and German (AI: 458) actually score *higher* than English (AI: 395) in this regard.

Regarding Szmrecsányi and Kortmann's [3] results for English (AI: 427; SI: 197), the present study's results (AI: 395; SI: 210) come very close. A Chi-squared test confirms that there is no statistically significant difference (AI: χ^2=1.25, df=1, p>0.5; SI: χ^2=0.42, df=1, p>0.5). Furthermore, Slovak (AI: 355; SI: 595) turned up similar scores as its close relative Czech, for which Szmrecsányi and Kortmann determined an AI of 334 and an SI of 683. This suggests that their method is indeed replicable.

However, German (AI: 458; SI: 595) deviates significantly from Szmrecsányi and Kortmann's results (AI: 436; SI: 301) with regard to syntheticity, as a Chi-squared test confirms (AI: χ^2=0.54, df=1, 0.5<p<0.1; SI: χ^2=57.04, df=1, p<0.001). One explanation might be that the texts from which the samples were drawn differed from each other. This is not entirely implausible – one of Szmrecsányi's studies showed, for example, that "variability in analyticity and syntheticity is endemic, surprisingly so, even among closely related dialects and varieties of the same

language" [1]. For clarification, it would be necessary to compare the data to Szmrecsányi and Kortmann's data set.

Figure 1 provides visualization of the results by means of Szmrecsányi and Kortmann's "typological space", where the *y*-axis "plots analyticity index scores while the [*x*]-axis indicates syntheticity index scores" [3]. Thus, a highly analytic language such as English will be in the top left corner (high analyticity, low syntheticity), whereas a synthetic language such as Slovak will be found in the lower right corner (low analyticity, high syntheticity). For comparison, the figure also contains Szmrecsányi and Kortmann's [3] results for English, German, and Czech (as gray data points).

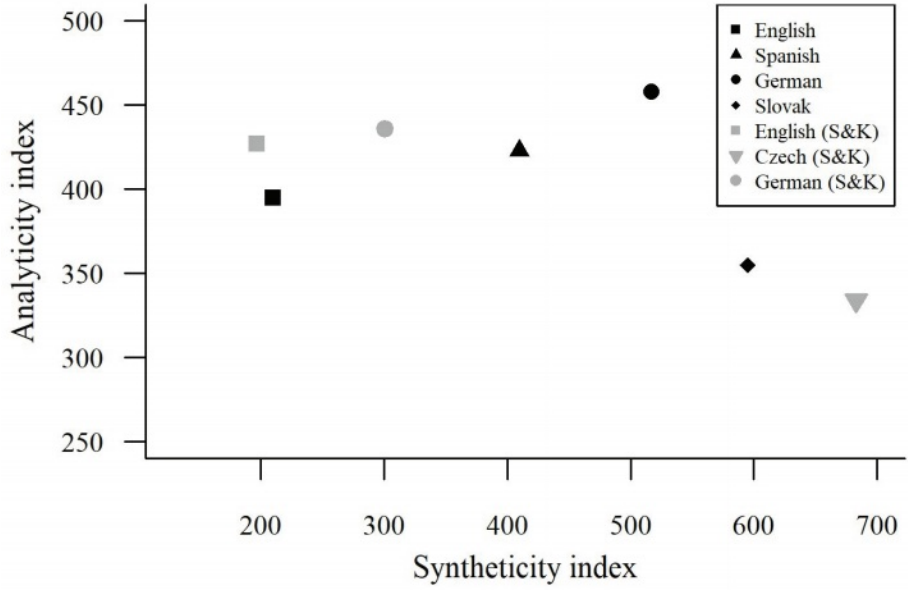


Fig. 1. Typological space: analyticity vs. syntheticity (in index points); index values from Szmrecsányi and Kortmann [3] in gray

The typological space facilitates comparison between languages. Figure 1 shows that in terms of syntheticity, the 'in-between' languages German and Spanish cover the middle ground between the 'extremes', English and Slovak. Another insight from the diagram is that analyticity and syntheticity are not necessarily exclusive categories or opposite poles of a one-dimensional continuum (as discussed in section 1): While English is indeed a "textbook example of a language that has developed from a synthetic language into an analytic one" [1], scoring high in analyticity and low in syntheticity, German and Spanish are not only more synthetic,

but also more analytic, a fact that is not easily appreciated without this kind of visualization.

4 CONCLUSION

The present study has demonstrated the feasibility of typologically profiling languages using Szmrecsányi and Kortmann's [3] methodology, that is, determining degrees of syntheticity and analyticity of languages based on naturalistic language data in the form of corpus samples. It was possible to confirm claims in the literature, such as English being "analytic to a very high degree" [4] and Slovak being "significantly inflecting" [5]. By calculating analyticity and syntheticity indices, such claims can now be substantiated with empirical evidence. It was also demonstrated that analyticity and syntheticity indices allow for a very precise comparison of languages.

It was furthermore shown that instead of an analytic-synthetic continuum, it is more appropriate to use a typological space consisting of two axes. This is because languages can be synthetic *and* analytic to varying degrees. It was also possible to replicate Szmrecsányi and Kortmann's [3] results regarding English and German, although German deviated considerably with regard to the syntheticity index. The present study has also highlighted the manifold possibilities in which corpus data can and should be employed in linguistics. Beyond exploring syntactic, lexical, and morphological phenomena, it can also be employed for the typological profiling of languages.

Finally, it must be noted that calculations of analyticity and syntheticity based on word/morpheme counts must always be taken with a grain of salt, because as usual, matters are more complex than they appear at first glance. As Schwegler notes, "many of the so-called analytic constructs […] have a considerably tighter morphological cohesion (i.e., are more synthetic) than the label analytic suggests" [8]. In this context, he mentions that "[Old French] *je* 'I' has a morphosyntactic and semantic profile which in many ways parallels that of its [Latin] ancestor *ego*, but […] in many ways differs fundamentally […] by having entered into a tighter relation with the verb" [8]. Similarly, the English pronoun *I* is in a tighter relation with the verb than its counterpart *ja* in Slovak, which is a pro-drop language like Latin.

Likewise, one might argue that there are varying degrees of syntheticity. Consider the Slovak locative case, which is marked on nouns with a suffix. A noun with a locative suffix would, therefore, add to the syntheticity index. However, locative-case nouns never appear without a preposition in Slovak, having lead some to even speak of a hybrid synthetic-analytic strategy [6]. However, the present study's approach simply takes words and morphemes and drops them into one of two bins – synthetic or analytic – instead of taking into account the context (e.g.

personal pronouns or prepositions) within which they are found. To address concerns about varying degrees of analyticity (e.g. pronoun-verb relation in English) and syntheticity (e.g. locative case with obligatory prepositions in Slovak), future research will have to take into account the context of words and how strongly they are 'attracted' to each other.

The present small-scale investigation, which is to be understood as a pilot study, holds considerable potential for future research. Calculating the analyticity and syntheticity indices of other languages, it is possible to further test and corroborate (or refute, for that matter) claims in the literature, e.g., about typological relatedness. For example, Czech and Slovak have been described as "two closely related languages" [16] – the present study could not only confirm this claim, but also determine just *how* closely they are related. From a diachronic perspective, syntheticity and analyticity indices can help trace the morphosyntactic evolution of languages [2]. As Schwegler notes, "many long-term diachronic changes cannot be grasped appropriately without the notions of analyticity and syntheticity" [8]. Finally, one issue that arises from using the Sketch Engine corpora is that they are composed of texts from the internet, which raises questions about their representativeness. Future research should, therefore, be based on data from more balanced corpora.

ACKNOWLEDGEMENTS

Thanks go to the Sketch Engine support team for helping me refine my initially rather cumbersome CQL expressions. I also highly appreciate the comments of two anonymous reviewers who helped improve my manuscript.

References

[1] Szmrecsányi, B. (2009). Typological parameters of intralingual variability: Grammatical analyticity versus syntheticity in varieties of English. Language Variation and Change 21(3), pages 1–35.

[2] Szmrecsányi, B. (2012). Analyticity and syntheticity in the history of English. In T. Nevalainen and E. C. Traugott (eds.), The Oxford Handbook of the History of English, pages 654–665. Oxford: Oxford UP.

[3] Szmrecsányi, B., and Kortmann, B. (2011). Typological profiling: learner Englishes versus indigenized L2 varieties of English. In J. Mukherjee and M. Hundt (eds.), Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap, pages 167–187. Amsterdam: Benjamins.

[4] Askedal, J. O. (2001). Conceptions of typological change. In M. Haspelmath (ed.), Language Typology and Language Universals: An International Handbook, vol. 2, pages 1624–1639. Berlin: De Gruyter Mouton.

[5] Dvonč, L., Horák, G., Miko, F., Mistrík, J., Oravec, J., Ružička, J., and Urbančok, M. (1966). Morfológia slovenského jazyka ['Morphology of the Slovak language']. (Ed.) J. Ružička. Bratislava: Slovenská akadémia Vied ['Slovak Academy of Sciences'].

[6] Oravec, J. (1980). Morfológia spisovnej slovenčiny [‘Morphology of written Slovak’]. Bratislava: Slovenské pedagogické nakladateľstvo [‘Slovak Publishing House of Educational Literature’].
[7] Schlegel, A. W. von (1818). Observations sur la langue et la littérature provençales [‘Observations on the provençal language and literature’]. Paris: Librairie grecque-latine-allemande.
[8] Schwegler, A. (1990). Analyticity and syntheticity: a diachronic perspective with special reference to Romance languages (Empirical Approaches to Language Typology 6). Berlin; New York: Mouton de Gruyter.
[9] Sapir, E. (1921). Language. New York: Harcourt, Brace and Co.
[10] Greenberg, J. H. (1960). A quantitative approach to the morphological typology of language. International Journal of American Linguistics 26(3), pages 178–194.
[11] V. Kasevič and S. E. Jachontov (eds.). (1982). Квантитативная типология языков Азии и Африки [‘A quantitative typology of Asian and African Languages’]. Leningrad: изда́тельство ленинградского университета [‘Leningrad University Press’].
[12] Real Academia Española. (2009). Nueva gramática de la lengua española [‘New grammar of the Spanish language’]. Vol. 1: Morfología, Sintaxis. Madrid: Espasa Libros, S. L. U.
[13] A. Wöllstein (ed.). (2016). Duden: Die Grammatik [‘Duden: The Grammar’]. 9th updated and completely revised edition. Berlin: Dudenverlag.
[14] Kilgarriff, A., Smrž, P., and Tugwell, D. (2004). The Sketch Engine. In G. Williams and S. Vessier (eds.), Proceedings of the eleventh EURALEX international congress (EURALEX 2004), pages 105–116. Lorient: Université de Bretagne-Sud, France.
[15] Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. (2014). The Sketch Engine: ten years on. Lexicography ASIALEX 1, pages 7–36.
[16] Nábělková, M. (2016). The Czech-Slovak Communicative and Dialect Continuum: With and Without a Border. In T. Kamusella, N. Motoki and C. Gibson (eds.), The Palgrave Handbook of Slavic Languages, Identities and Borders, pages 140–184. London: Palgrave Macmillan.
[17] The Document Foundation. (2021). LibreOffice Calc v. 7.1.2. Available at: https://www.libreoffice.org/discover/calc/.
[18] Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). A Comprehensive Grammar of the English Language. New York: Longman.
[19] Gallmann, P. (2016). Die flektierbaren Wortarten [‘Inflective word classes’]. In Duden: Die Grammatik [‘Duden: The Grammar’], pages 149–385. 9th updated and revised edition. Berlin: Dudenverlag.
[20] Ledgeway, A. (2017). Syntheticity and Analyticity. In A. Dufter and E. Stark (eds.), Manual of Romance Morphosyntax and Syntax, pages 839–886. Berlin: De Gruyter Mouton.

sciendo
DOI 10.2478/jazcas-2021-0033

ACQUIRING WORD ORDER IN SLOVAK AS A FOREIGN LANGUAGE: COMPARISON OF SLAVIC AND NON-SLAVIC LEARNERS UTILIZING CORPUS DATA

MARTINA IVANOVÁ[1] – MIROSLAVA KYSEĽOVÁ[1,2] – ANNA GÁLISOVÁ[3]
[1] Faculty of Arts, University of Prešov, Prešov, Slovakia
[2] Faculty of Philology, Jagiellonian University, Kraków, Poland
[3] Faculty of Arts, Matej Bel University, Banská Bystrica, Slovakia

IVANOVÁ, Martina – KYSEĽOVÁ, Miroslava – GÁLISOVÁ, Anna: Acquiring word order in Slovak as a foreign language: Comparison of Slavic and Non-Slavic learners utilizing corpus data. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 353 – 370.

Abstract: The paper deals with the acquisition of Slovak word order in written texts of students of Slovak as a foreign language. Its attention is focused on identifying the correct and incorrect placement of enclitic components, and their erroneous usage is analysed with respect to different investigated variables (types of enclitic components, types of syntactic construction, distance from lexical/syntactic anchor, and realization in pre- or post-verbal position). The paper also pays attention to the error rate regarding individual proficiency levels of students, and error distribution in two language groups, Slavic and Non-Slavic learners, is compared.

Keywords: word order, enclitics, error analysis, syntactic complexity, Slavic learners, Non-Slavic learners, acquisition stages, interlanguage

1 INTRODUCTORY REMARKS

Among morpho-syntactic phenomena, one of the most problematic challenges for students of Slovak as a foreign language is acquiring word order. The main reason is that Slovak word order is formed on the borderline of three major principles, i.e., functional sentence perspective, prosody, and grammar, which represent independent factors determining the linear order of a sentence, yet they sometimes interfere with each other. Since linearization of Slovak sentence structure is not determined by the grammatical functions of sentence components (except for attributes within noun phrases) and its major function is to express information structure (see [1]), it is characteristic of relative flexibility. On the other hand, word order flexibility is highly restricted with respect to position of attribute phrases and clitics due to grammatical and prosodic rules which govern their placement. Clitics, especially, represent one of the most specific and intricate phenomena within the word order of many languages. Even the languages with most similarities differ in

clitic placement, as pointed out by Uhlířová regarding Czech, Slovak, and Polish [2, p. 82]. This fact also opens up space for their erroneous usage in texts produced by highly proficient speakers.

The present paper is aimed at investigating acquisition of enclitics ordering by foreigners learning Slovak as L2. Based on performance analysis [3, p. 73], the aim is to map accuracy in the placement of enclitics with respect to the level of language acquisition (lower proficiency versus higher proficiency level) and the affiliation of the learners' mother tongue(s) to a language group (Slavic versus non-Slavic language family). To get a closer picture of Slovak word order acquisition, we compiled our own corpus of written texts of students learning Slovak as a foreign language at different proficiency levels, and we supplied each enclitic component present in the texts with annotation tags reflecting different variables. By measuring the error distribution and relating the statistical values with specific features of the texts (syntactic complexity), our aim was to specify how fully learners of the two language groups acquire the principles of enclitic ordering at different stages of their interlanguage.

However, determining the extent to which learners use a certain language feature accurately presupposes identification of an error and its distinction from correct use. Within the Slovak context, one particular circumstance that hinders identification of erroneous or inappropriate word orders is the absence of theoretical and practical investigation into word order, which would show preferential patterns of word order in Slovak as L1. The only work on this topic in the Slovak context was published in 1966 by J. Mistrík [4]; other works concerning Slovak word order focus mostly on syntagmatic word order (cf. [5]) and are not based on corpus data. The description of Slovak enclitics and their linear ordering within the Slavic context can be found in Frank & King [6] and Beličová & Uhlířová [7]. The situation in Slovak is largely at odds with that of Czech in which word order behaviour of clitics attracted significant attention both in investigation of Czech as L1 (cf. [8] for an overview) as well as L2 (cf. [9]).

The structure of the present paper is as follows: based on the theoretical literature available, in Section 2, we present a short definition and classification of enclitics in Slovak. In Section 3, we describe our samples and methods for annotating these enclitics. In Section 4, statistical results are presented, and Section 5 is devoted to discussion of the results and the conclusion.

2 CLITICS IN SLOVAK WORD ORDER

2.1 Definition of clitics

Prosody relates to the word order realization of phonologically non-independent elements devoid of word stress called clitics, which cannot be realized freely, i.e., in various sentence positions depending on the pragmatic and discourse function, but

their position within the sentence structure is determined phonologically. Slovak belongs to those languages which follow Wackernagel's Law and its clitic elements belong to the category of second-position clitics (2P)[1] [10], which are typical of having "dual citizenship". Within the sentence, they follow an initial element called prosodic host (anchor), a clause-initial unit, usually the first sentence constituent. However, morphologically, lexically, or syntactically they belong to the governor, most typically a verb position not conditioned prosodically within the sentence. As the prosodic host of the clitic component and its governor do not necessarily correspond, it may lead to emergence of constituent discontinuity (cf. [1]).

Phonologically, 2P clitics are enclitics, however, they may be procliticized under certain circumstances. It happens in complex sentences with a matrix clause containing a clitic item which is realized discontinuously, being disrupted from the initial sentence component by an interposed subordinated clause. If the clitic component is realized after the interposed clause (after a pause), phonologically, it is procliticized to the following sentence constituent, e.g., *Samozrejme aj to, čo je na tanieri,* ***ma*** *inšpiruje.* 'Of course, everything on the plate inspires me, too.' (Omnia Slovaca III).

2.2 Classification of clitics in Slovak

2.2.1 Constant and inconstant clitics

Prosodic deficiency (the absence of word stress) is not always considered as the defining feature of clitic components. In many theoretical works, constant and inconstant clitic components are differentiated, the former labelled enclitics tantum and the latter as volatile enclitics (cf. [4] for Slovak). The following characteristics can be stipulated for those two groups:

(i) Enclitics tantum, or pure sentential clitics (cf. [11] for the term), can be defined as prosodically deficient unstressable elements that are unstressed independently of the context in which they are realized, thus, they are unable to be focused and cannot be moved to initial position. According to Junghanns [12], they can be labelled as lexical clitics, as the clitic status represents an inherent part of their lexical "equipment".

[1] Despite the fact that Slovak clitics are defined as second position clitics, there are many deviations from that rule. Ambiguity of clitic placement holds especially true for two structure types: (i) for compound and complex sentences with certain complementizers, e.g., after *ale* ('but'), the enclitic component can either occupy the position immediately after the complementizer: *Vydala sa za nejakého Bergera, ale* ***sa*** *s ním rozviedla.* 'She married a certain Berger but she divorced him.' (Omnia Slovaca III), or after the first sentence constituent: *Cítila jeho dych za chrbtom, ale neobrátila* ***sa***. 'She felt his breath behind her back but she did not turn around.' (Omnia Slovaca III); (ii) for sentences with a multi-constituent thematic part: the enclitic component can either occupy the second position (after the complementizer): *Teraz prišiel trest za to, že* ***sa*** *kedysi vzdala svojho syna.* 'Now came the punishment for her giving up her son.' (Omnia Slovaca III), or it is realised after the first thematic item: *S. Markovič spomína, že kedysi* ***sa*** *náklad nosil hore síce ťažšie, ale oveľa romantickejšie.* 'S. Markovič recalls that carrying the load up used to be more difficult, though much more romantic.' (Omnia Slovaca III).

(ii) Volatile/inconstant enclitics or semi-clitics (cf. [11] for the term) can be defined as prosodically unstressed elements that can have phonological autonomy under certain contextual conditions. In Junghanns [12], they are labelled as phonological clitics due to the fact that their clitic status is formed "in the phonological part of the [sic] grammar usage".

However, the boundary between clitics and non-clitics is often blurred. This is especially the case of inconstant/volatile enclitics. As Hana [1, p. 74] points out, enumerating the exact set of clitics is far from trivial and probably impossible. In our approach, clitic status is ascribed to those monosyllabic auxiliary and non-auxiliary and bi-syllabic auxiliary components[2] which conform to the property specified by Hana ([1]): [1P–Cl] A word between 1P and a clitic is a clitic.[3] The dataset of investigated clitic components will be specified in the following section. Due to low frequency in students' texts, we also decided to omit clitic conjunctions and particles from our investigation.

2.2.2 Verbal and argument clitics

Enclitics can be further divided into two categories depending on the possibility to independently fulfil syntactic functions within the sentence. Dependent morphological enclitics relying on their lexical governor (the verb) and functioning as exponents of grammatical categories (tense, mood, voice, person, number) can be labelled as verbal clitics. In Slovak, the following verbal enclitic components can be differentiated: (i) conditional enclitic tantum: exponent of mood (the clitic component marking conditional mood), (ii) auxiliary enclitics tantum: exponents of person and number (auxiliary components marking person and number in l-participle forms of the preterite and antepreterite *som, si, sme, ste*), (iii) reflexive enclitic tantum: exponent of voice (the reflexive clitic *sa* marking passive voice within reflexive deagentive constructions).

On the other hand, independent lexical enclitics capable of fulfilling syntactic functions within the sentence as verbal complements (with the verb as their syntactic governor) can be labelled as argument clitics. Argument clitics are represented by weak or short pronominal forms, coding both direct and indirect objects or adverbials. Among argument clitics, the following subgroups can be singled out: (i) non-prepositional personal enclitics tantum: short forms of personal pronouns which stand in opposition to long accented forms (*ma – mňa* 'me'*, ťa – teba* 'you'*, ho – jeho* 'him'*, mi – mne* 'me'*, ti – tebe* 'you'*, mu – jemu* 'him'), (ii) non-prepositional personal volatile enclitics:

[2] Due to unclear status in theoretical studies, we decided to exclude bi- and tri-syllabic prepositional-pronominal forms from our dataset, e.g., *s nami* 'with us'*, pre mňa* 'for me'*, na neho* 'on him'.

[3] Hana [1, pp. 75–76] uses two other criteria for clitic delimitation: "Clitics cannot occur in isolation, e.g., as an answer to a question.", "Clitics cannot occur sentence-finally". As inconstant clitics are also included in our dataset, we do not apply these criteria.

unparalleled forms of personal pronouns which can be used either as enclitics or as accented full forms (*ju* 'her', *nás* 'us', *vás* 'you', *ich* 'them', *jej* 'her', *nám* 'us', *vám* 'you', *im* 'them'), (iii) demonstrative volatile enclitics: forms of demonstrative pronouns which can occupy the initial position in the stressed form or under certain conditions they become prosodically dependent and behave like enclitic elements (*to* 'that', *tu* 'here', *tam* 'there', *tak* 'so', etc.), (iv) prepositional volatile enclitics: forms of personal and demonstrative pronouns (*k nám* 'to us', *k vám* 'to you', *k nej* 'to her', *s ním* 'with him', *s ňou* 'with her', *s tým* 'with that', etc.), (v) reflexive enclitic tantum: reflexive pronouns *sa, si* functioning as weak, unstressed forms of the longer forms *seba, sebe*. The clitic status of the copular *byť* 'to be' is disputable and there is disagreement as to the clitic nature of the copular 'be' in Slavic languages. The opinions on the enclitic status of *'be'* forms can be classified as follows:

– enclitic status is assigned only to the auxiliary *be*-forms (e.g., [8], [6]);
– only the auxiliary *byť* in the present tense within passive constructions is labelled as an enclitic component (e.g., [11]);
– only auxiliary forms of the past conditional and antepreterite (*byl/a* for Czech) and present forms of the non-auxiliary *být* can acquire clitic status (e.g., [13]).

As can be seen, the deciding criterion for assessment of enclitic status to a component is associated with the degree of grammaticalization (auxiliaries as the most grammaticalized elements). However, according to Palková [14], it is a common process that monosyllabic elements often lose stress and became part of the prosodic tact of the neighbouring word. She states that it is a matter of rhythm, not grammatical status. If Palková's assumption is right, then the nature of *be*-forms is not determined by its grammatical status (auxiliary vs. copular vs. full lexical), but by contextual distribution. In that sense, the copular *byť* can also be described as a volatile enclitic element. In our data, enclitic status is assigned to: (i) monosyllabic forms of *byť* (auxiliary, copular, full lexical) realized in the second position: **Teraz Kabula je šťastná a žije v Poľsku. – Teraz je Kabula šťastná a žije v Poľsku.* 'Kabula is now happy and lives in Poland.' = copular *byť* (Polish, A2), *Všetko v tom meste jej pripomínalo Marka, ktorý je teraz niekde nad ňou* 'Everything in the town reminded her of Mark who is now somewhere above her.' = full lexical *byť* (Serbian, B2), (ii) disyllabic forms of *byť* (auxiliary elements in past conditional, antepreterite and periphrastic passive) realized in the second position: **Víťaz vyberal cieľ charitatívny, na ktorý dávaná bola cena. – Víťaz vyberal cieľ charitatívny, na ktorý bola dávaná cena.* 'The winner would select a charitable cause for the prize to be awarded to.' (Polish, A2).

Enclitic components can aggregate into clitic clusters comprising (hypothetically) 2 – 7 components. The internal organization of clitic clusters in Slovak can be described as follows: BY > AUX > REFL > NON-ARG. DAT > ARG. DAT > ACC > GEN > CONJ. However, in real communication, such extensive clusters are rather rare. In our data, clitic clusters usually comprise 2 to 3 members.

Mono-syllabic forms of the verb *byť* 'to be' (independent of their lexical status) can occur as part of clitic clusters between 1P constituent and another clitic, e.g., *Skoro je mi ťa ľúto.* 'I almost feel sorrow for you.' (Omnia Slovaca III). Bi-syllabic forms do not show similar behaviour, compare: **Skoro bolo mi ťa ľúto., Skoro mi ťa bolo ľúto.* That's why monosyllabic forms of the verb *byť* are treated as inconstant enclitics and form part of our dataset.

3 METHODOLOGICAL ASPECTS OF RESEARCH

3.1 Data description

To conduct our investigation of word order errors in texts written by foreigners learning Slovak, we compiled our own corpus of written texts. The data come from a pre-pilot version of the Corpus of Texts of Students Learning Slovak as a Foreign Language (errkorp-0.1) [15] which is under development. In its current state, the corpus comprises 12 733 tokens and 10 428 words. As the volume of the given corpus with respect to the amount of word order errors was not sufficient, we completed it with our own texts.[4] All sentences with enclitic components were transcribed into Excel and were assigned annotation tags reflecting the investigated variables (see 4.2). Overall, we analysed 81 texts, of which 43 texts come from students of Slovak with a Slavic mother tongue and 38 texts were produced by students with a non-Slavic mother tongue.[5] The texts were divided by proficiency level into two categories: 54 texts at the lower proficiency levels A1 – B1 and 27 texts at the higher proficiency levels B2 – C1 (according to CEFR). The aim was to obtain approximately 50 errors of enclitic placement in both language categories of texts at both investigated levels (A1 – B1 and B2 – C1), which is around 200 errors in total, which we considered to be the minimum amount for our analysis purposes. Overall, 1305 sentences with clitic components were analysed out of which 217 contained errors in enclitic usage and in 1089 enclitics were used correctly. The entire database is published online at zenodo.org (cf. [16]).

3.2 Error annotation

In errkorp-0.1, word order errors are divided into two categories from a predefined error taxonomy: the error tag ORDER is used for errors concerning enclitic components and attributive phrases and the error tag THEME is used for errors concerning functional sentence perspective.

[4] The non-corpus texts were obtained from lecturers of Slovak as a foreign language and were collected from non-native speakers of Slovak attending university language courses abroad. The texts were produced during different types of situations, i.e., in class, as homework and in examinations, and were handed in either in electronic form or as manuscripts.

[5] The data comprise: (i) texts of students from all three language groups within Slavic languages (West Slavic – Polish, South Slavic – Serbian and East Slavic – Ukrainian), (ii) texts of students with a non-Slavic Indo-European mother tongue, with a majority of Germanic (mostly English, German) and Romance (mostly Italian) languages, (iii) texts of students with non-Indo-European mother tongues, with a majority of Finno-Ugric (Hungarian) and Sino-Tibetan (Chinese) mother tongues.

As we need to analyse highly specialized language phenomena (word order of enclitic components), we decided to annotate corpus data manually with respect to the additional variables under investigation. We did not use any commonly used programmes for the purpose of compiling a corpus as the common options that those programmes offer (like tokenization, tagging, parsing, etc.) are not relevant to our investigation.

The texts in our database were annotated using two annotators independently, the annotations were later compared to eliminate subjective evaluation of errors. Agreement in annotation solutions achieved by the annotators was evaluated using the metric κ (kappa, cf. [17]) which is used as a standard measure instrument for inter-annotation agreement. It is calculated as:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where P(A) is observed agreement between the annotators, and P(E) is the expected agreement, i.e., the probability that the annotators agree by chance. The calculated interval oscillates between (0.1) where $\kappa = 1$ means perfect agreement and $\kappa = 0$ agreement equal to chance (cf. [18]). As a result, the reliability of the annotation has been proved as the calculations showed that = 0.92 which can be interpreted as nearly perfect agreement.[6]

3.3 Annotation parametres

To investigate acquisition of clitic ordering, we recorded the presence of all enclitic components in the 81 analysed texts, both correctly and incorrectly used. During text annotation, we took following parameters into account.

Analysed parameter	**Types**
Type of component	**Reflexive enclitics** R = reflexive component **Verbal enclitics** G = auxiliary *byť* (separate grammatical morphemes coding person and number in preterite and antepreterite forms) K = conditional morpheme **Argument enclitics** P = short form of personal pronoun PP = prepositional pronoun D = monosyllabic demonstrative pronoun **Be enclitics** S = non-auxiliary *byť* **Combination of enclitics in a row** KT = clitic cluster

[6] 1082 cases of enclitic usage were rated as Correct by both annotators, 204 cases were rated as Errors by both annotators, so P(A) = 0.98. Annotator A rated 1088 cases as Correct, Annotator B 1095 cases; Annotator A rated 217 cases as Errors, Annotator B 210 cases, so, P(E) = 0.73. $\kappa = (0.98 - 0.73) / (1 - 0.73) = 0.92$.

Analysed parameter	Types
Correctness of usage	C = correct
	E = error
Type of syntactic construction	JV = simple sentence (including the first main clause in a compound sentence, initial main clause in a complex sentence)
	PS = second main clause in a compound sentence
	HSH = postponed main clause of a complex sentence
	HSV = subordinate clause of a complex sentence
	IK = reflexive component in non-finite construction, mostly infinitive
Correct proximity to lexical/syntactic host	0 = zero distance 1 = 1 phrase between an enclitic and the host 2 = 2 phrases between an enclitic and the host...
Correct position in relation to lexical/syntactic host	preV = preverbal position postV = postverbal position

Tab. 1. Parameters analysed with respect to the prosodic factor

4 RESULTS

In this section, the outcomes of error distribution measuring with respect to the stages of acquisition and the investigated variables will be presented. The aim is 1) to quantify and compare the ratio of correct and erroneous usage of enclitic components, 2) to describe the relationship between syntactic complexity and erroneous usage of enclitics, 3) to identify similarities in distribution of errors with respect to the investigated variables at early and intermediate (A1 – B1) and advanced (B2 – C1) stages of acquisition.

4.1 Correct and erroneous usage in the texts

The ratio of correct and erroneous enclitic placement in texts by the Slavic and the Non-Slavic groups is provided in the following figure.[7]

[7] At elementary and lower-intermediate levels, 307 sentences (26 texts) were annotated within the Slavic group and 223 sentences (28 texts) were annotated within the Non-Slavic group. At the upper-intermediate and advanced levels, 502 sentences (17 texts) were annotated within the Slavic group and 273 sentences (10 texts) were annotated within the Non-Slavic group.

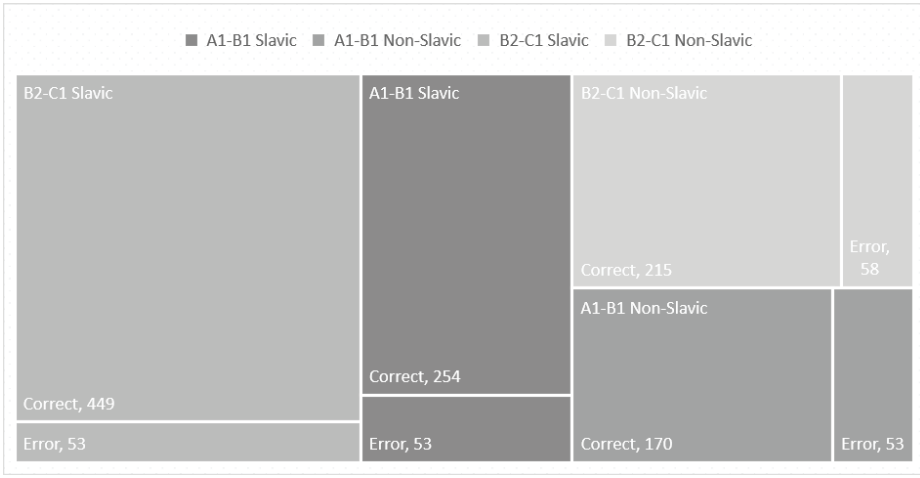


Fig. 1. The ratio of correct and erroneous usage in Slavic and Non-Slavic texts

The data show that, at both the lower and the higher-proficiency levels, Non-Slavic speakers produce more errors concerning enclitic elements than Slavic students: at A1-B1 levels more than every fourth enclitic component (4.2) is used erroneously in the Non-Slavic texts compared to almost every sixth erroneous component (5.8) in the Slavic texts; at B2 – C1 levels more than every fourth enclitic component (4.7) is used erroneously in the Non-Slavic texts as opposed to more than every ninth incorrectly placed component (9.5) in the Slavic texts.

4.2 Syntactic complexity of texts

The results presented in the previous section (4.1) should be elaborated against the background of phenomenon labelled as syntactic complexity, which is considered an indicator of overall level of L2 proficiency (cf. [19]). In numerous studies concerning second language acquisition, one aspect that syntactic complexity has been approached from was represented by subordination ratio, i.e., the ratio of embedded syntactic structures deemed to be developmentally or cognitively complex (e.g., sub-clauses) (cf. [20]).[8] Syntactic complexity in L2 is thought to expand from coordination to subordination and then to phrasal elaboration, as learners gain proficiency (cf. [27]). At beginner and low-intermediate proficiency levels, syntactic

[8] It has been claimed that, at initial levels, L2 learners start producing simple clauses organized around finite verbs (stage 1, cf. [21]), subsequent developmental stages involving the shift from simple clauses to the use of clause linking (stage 2), at advanced levels, the development involves noun and verb phrase elaboration, when information previously encoded as a clause is embedded (stage 3, cf. [22], [23]), and the highest stage is connected with sub-clausal complexification at the phrasal level, which is supposed to be characteristic of academic discourse and written prose (cf. [24], [25], [26]).

growth may show an increase in coordination (e.g., [28], [29]) and upper-intermediate levels are thought to display an increase in subordinate structures.

To verify theoretical assumptions on L2 development reflected in growing syntactic complexity, we calculated the ratio of simple clauses, compound and complex sentences in our sample texts for both language groups (see the following table).

		JV	**PS**	**HSV + HSH**	**IS**
	Non-Slavic	61%	16%	21%	2%
A1_B1	Slavic	49%	18%	30%	3%
	Non-Slavic	51%	11%	37%	1%
B2_C1	Slavic	49%	11%	38%	2%

Tab. 2. Syntactic complexity in the Non-Slavic and Slavic texts at different proficiency level

The data from our investigation show three major tendencies: (a) a drop in the ratio of simple clauses at higher proficiency levels in the Non-Slavic texts[9], (b) a drop in the ratio of compound sentences at higher proficiency levels in both the Slavic and Non-Slavic texts, (c) and, at the same time, a rise in the ratio of complex sentences at higher proficiency levels in both the Slavic and Non-Slavic texts.

We also calculated the distribution of compound sentences at individual lower proficiency levels, as can be seen in the following figure.

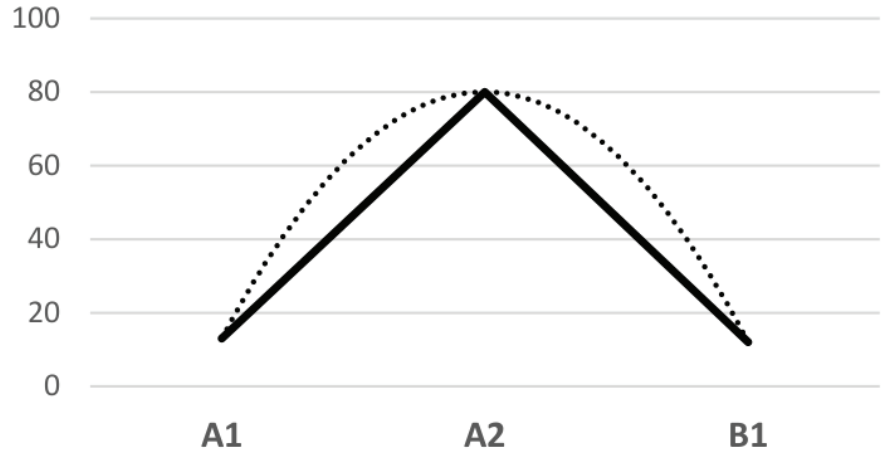


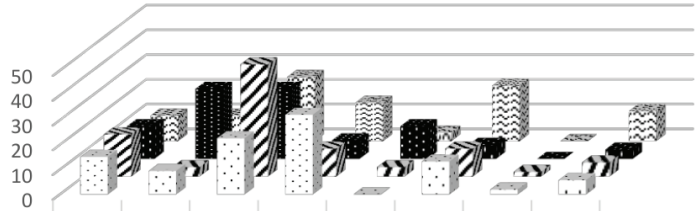
Fig. 2. Frequency distribution of compound sentences at individual levels

[9] Slavic texts display a relatively low number of simple clauses even at lower proficiency levels when compared with Non-Slavic texts.

The outcomes appear to confirm the developmental "omega-shaped" pattern suggested by Wolfe-Quintero et al. [28], which is to say a decrease in coordination at higher proficiency levels in favour of subordination. At the same time, our data correspond with the results of those studies which have indicated a rise in subordination rates at intermediate levels (e.g., [27]). Our data show a peak in coordination usage at A2 level and gradual decline in its usage beginning at B1 level.

4.3 Frequency distribution of errors with respect to enclitic type

Frequency distribution of errors with respect to enclitic type in the texts of Slavic and Non-Slavic students shows certain tendencies which can be visualized through the following figure.



	COP	DEM	REF	PRON	PP	AUX	COND	CLUST
Slavic A1-B1	15,1	9,4	22,6	32,1	0	13,2	1,9	5,7
Non-Slavic A1-B1	17	3,8	45,3	11,3	3,8	11,3	1,9	5,7
Slavic B2-C1	13,2	28,3	28,3	7,5	13,2	5,7	0	3,8
Non-Slavic B2-C1	10,3	10,3	25,9	15,5	3,4	22,4	0	12,1

Fig. 3. Error distribution in the texts of Slavic and Non-Slavic students

By comparing texts by Slavic and Non-Slavic speakers at A1 – B1 level, the most striking differences in frequency distribution can be grouped into two main categories:

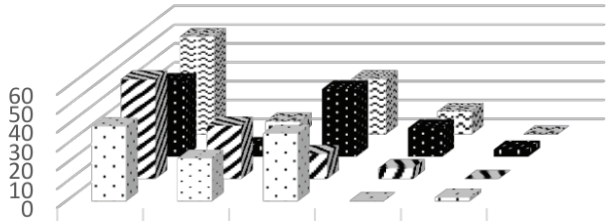
(a) argument pronominal and demonstrative clitic errors are more frequent in the texts of the Slavic students (for the Slavic A1 – B1 texts, the distribution of D, P and PP errors is 41.5%; for the Non-Slavic A1 – B1 texts it is only 18.9%), e.g., *Preto bolo* ***im*** *veľmi ľúto a mysleli, že vždy takto bude.* 'That's why they were very sorry, and they thought that it would always be that way.' (A2, Polish);

(b) verbal clitic errors (including R errors and G errors) are much more frequent in the texts of Non-Slavic students (35.8% in Slavic texts and 56.6% in the Non-Slavic texts), e.g., *A* ***sme*** *našli nejaké zaujímavé veci o čínskej a slovenskej kuchyni.* 'And we found out some interesting facts about Chinese and Slovak cuisine.' (B1, Chinese).

At higher proficiency levels, two tendencies are observed: (a) in Slavic texts, the distribution of P errors dramatically decreases in favour of an increasing number of PP and D errors, i.e., Slavic students produce more errors concerning volatile PP and D enclitics whereas the correct placement of P enclitics tantum is relatively acquired in their interlanguage (there is no striking difference between constant and volatile argument clitics in the texts of Non-Slavic students); (b) the ratio of errors concerning verbal clitics is still higher in the texts of Non-Slavic students (48.3%) when compared to Slavic students (34%), however, the difference is not so striking when compared to lower proficiency levels.

4.4 Frequency distribution of errors with respect to type of syntactic structure

The following figure features frequency distribution of errors with respect to type of syntactic construction.



	JV	PS	HSV	HSH	IS
Slavic A1-B1	39,6	22,6	35,8	0	1,9
Non-Slavic A1-B1	52,8	28,3	13,2	5,7	0
Slavic B2-C1	39,6	5,7	35,8	15,1	3,8
Non-Slavic B2-C1	51,7	6,9	29,3	12,1	0

Fig. 4. Error distribution with respect to syntactic structure

The data from Figure 4 are coherent with the results concerning syntactic complexity of texts at individual proficiency levels (i.e., the error rate in individual sentence types corresponds to the overall distribution of sentence types, cf. Table 2). At lower and intermediate proficiency levels, these findings emerged from the data:

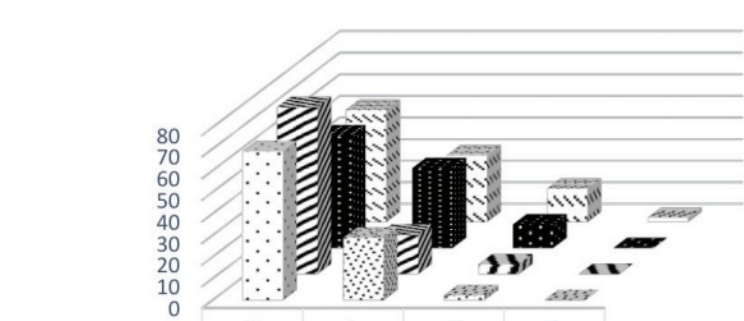
(a) the Slavic texts show the highest erroneous usage with respect to complex sentences, e.g., *Myslim si že pokoriš* ***ho***. 'I think that you will break it.' (A2, Ukrainian);

(b) in the Non-Slavic texts most errors occur in simple sentences, e.g. *Toto zaujalo detektiva, on myslel* ***si*** *že budu nasledovať ešte zločiný.* 'This interested the detective; he thought that more crimes would ensue.' (B1, German).

At upper-intermediate and advanced levels, the number of errors occurring in complex sentences in the Non-Slavic texts rises considerably (18.9% vs. 41.4%) which can be associated with the increasing syntactic complexity of the Non-Slavic texts. Both in the Slavic and Non-Slavic texts, there is a striking drop in the number of errors occurring in compound sentences, which can be explained by the "omega-shaped" pattern in the distribution of compound sentences (cf. 4.2). The number of errors occurring in simple sentences remains the same at lower and higher proficiency levels both in the Slavic and Non-Slavic texts.

4.5 Frequency distribution of errors with respect to distance from lexical/ syntactic host

The same comparison can be made with respect to the distance of enclitic components from their lexical/syntactic hosts as shown in Figure 5.



	0	1	2	3
Slavic A1-B1	68,9	28,9	2,2	0
Non-Slavic A1-B1	77,3	18,2	4,5	0
Slavic B2-C1	52,2	37	10,9	0
Non-Slavic B2-C1	51,9	30,8	15,4	1,9

Fig. 5. Error distribution with respect to the distance from host component

The data from Figure 5 show that in both the Slavic and Non-Slavic speakers' texts most errors relate to enclitics with 0- and 1-distance position. At elementary and lower-intermediate levels, the following tendencies can be observed:

(a) the errors occur in 0-distance in the Non-Slavic texts more often than in the Slavic texts, e.g., *Študenti a dôležití hostia sa zúčastnili a* ***sme*** *pozerali deväť krátke filmy.* 'Students and important guests took part, and we watched nine short films.' (A2, Italian);

(b) higher error distribution in 1-distance is observed in the Slavic texts, e.g., *Mne velmi* ***sa*** *páči cestovať, preto že najlepši deň pre mňa to je deň kedy začinaje cestovanije.* 'I like travelling very much, and because of this the best day for me is the day when a trip begins.' (A2, Ukrainian).

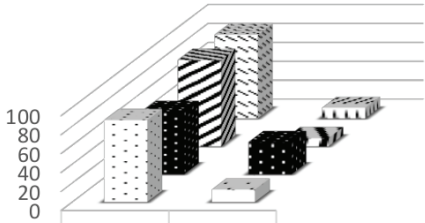
It can be linked to the fact that Slavic learners at these proficiency levels produce more complex syntactic structures (cf. 4.2) with enclitic components often put in more distant positions from their lexical/syntactic hosts which may cause a higher occurrence of 1-distance errors.

At upper-intermediate and advanced levels, there is a distinct decrease in erroneous placement of enclitics in 0-distance and an evident increase in 1-distance and 2-distance placement in both the Slavic and Non-Slavic texts. It seems that, in their interlanguage development, both Slavic and Non-Slavic learners are able to place the enclitic component in the proper position more often when it is adjacent to its lexical/syntactic host and the higher erroneous usage of enclitics relates to their distant, non-adjacent positioning.

At the same time, at higher proficiency levels, there is no striking difference between the Slavic and Non-Slavic texts, which corresponds to our findings according to which the syntactic complexity of the Non-Slavic texts is approaching the Slavic texts regarding interlanguage development, which is reflected in converging distribution of errors in the texts.

4.6 Frequency distribution of errors with respect to pre- and post-verbal position

Finally, the erroneous usage of enclitics is connected to the ability of learners to shift enclitics into pre-verbal position or to place them in post-verbal position. The results are presented in Figure 6.



	Pre-verbal	Post-verbal
Slavic A1-B1	86,7	13,3
Non-Slavic A1-B1	65,9	34,1
Slavic B2-C1	91,3	8,7
Non-Slavic B2-C1	88,5	11,5

Fig. 6. Error distribution with respect to pre-verbal or post-verbal position

As the data from Figure 6 show, the erroneous usage of enclitics is connected mostly with the ability to put them into preverbal position in the Slavic texts whereas the ratio of erroneous distribution in post- and pre-verbal position is more balanced in

the texts of Non-Slavic students, however, it holds true only for lower and intermediate proficiency levels. The higher frequency of erroneous usage in post-verbal position in the Non-Slavic texts can be caused by higher frequency of compound sentences in the texts by Non-Slavic students in which the proclitic conjunction, which requires postponing enclitic elements into post-verbal position, is often employed, e.g., *Niekedy vybuchne, stále odbieha od problémov a* ***si*** *nenechá čas na ich riešenie.* 'Sometimes he loses control; he's always ignoring his problems, not leaving himself the time to solve them.' (C1, English). At upper-intermediate and advanced levels, the distribution of errors is even more attracted to pre-verbal position both in the Slavic and Non-Slavic texts, e.g., *Povedala mi, že spoznali* ***sa*** *na diskotéke.* 'She told me that they had met at a club.' (C1, Polish). It can be explained by a decrease in compound sentence structures, which usually motivates post-verbal position of enclitic components.

5 DISCUSSION AND CONCLUSIONS

Based on the data (see Figure 1), at first glance, it seems that Slavic speakers at higher proficiency levels show progress (producing fewer errors), and conversely, non-Slavic speakers' word order acquisition more or less stagnates, i.e., almost every fifth (4.7) enclitic is still placed inaccurately in their texts (when compared to the lower proficiency level, where erroneous distribution concerns every fourth component (4.2)). However, investigation into development of syntactic complexity sheds new light on the issue. As indicated in Section 4.2, syntactic complexity varies in our sample texts with respect to proficiency levels, achieving a higher degree at upper-intermediate and advanced levels, which holds true for both Slavic and Non-Slavic text groups. Despite more syntactically complex nature of their texts, Slavic learners produce lower – and Non-Slavic learners produce comparable – amounts of errors. In other words, increasing syntactic complexity does not result in a greater number of enclitic errors. This finding leads us to the conclusion that acquisition of word order goes hand in hand with higher L2 proficiency in both language groups.[10]

Against the background of different syntactic complexity of the Slavic and Non-Slavic texts at lower levels of proficiency, the uneven frequency distribution of errors concerning pronominal enclitics can also be explained. As shown in Section 4.3 (see Figure 3), the ratio of argument enclitic errors is higher in Slavic speakers' texts than in the Non-Slavic ones. However, not only errors, but also the overall distribution of argument enclitics is higher in the texts of Slavic students. The more frequent usage of pronominal argument enclitics may be related directly to the higher ratio of complex sentences, since object pronouns (for 3rd person) require an antecedent to which the

[10] Worth noting is the fact that the degree of syntactic complexity in the Slavic texts is higher than in the Non-Slavic texts even at elementary and lower-intermediate levels. This issue goes beyond the scope of our study but opens up space for further comparative research into acquisition of Slovak by Slavic and Non-Slavic learners at early stages of their L2 development.

form is referring, so that this type of construction often requires more than one sentence, as concepts are developed across longer strings of language. At initial stages, the use of repetition as an avoidance strategy has proved dominant (cf. [30]) which seems to be the case of the Non-Slavic students' texts in our sample. However, the issue was not explored in more detail in our study and is open to further research.

At the upper-intermediate and advanced levels there is also a slight decrease in the number of copular enclitic errors both in the Slavic and in Non-Slavic texts (see Figure 3). At the elementary and lower-intermediate level, the copular *byť* 'to be' belongs to those basic verbs which are acquired at the early stage and is typical of high dominance in terms of frequency (cf. [31]). In other words, the higher frequency of copulas in general is reflected in higher distribution of copular errors at initial stages of L2 development.[11]

Finally, the data presented in Section 6 (cf. Figure 6) also point out that, in interlanguage development, there is a cognitive barrier blocking the correct preverbal position of enclitic components. Two possible explanations are at hand with respect to this phenomenon:

(i) As to verbal enclitics: the components with grammatical function are typically realized at the right periphery of the verb which corresponds with the investigation of affix ordering in Slavic languages (cf. [32]): affixes are realized in the order: prefix-basis-suffix-thematic marker-grammatical morphemes (this can be an explanation for the fact that placement of verbal enclitics with grammatical function is attracted to the right periphery of the verb in the texts, irrespectively of prosodic conditions in syntactic structures).

(ii) As to argument enclitics: as argument clitics usually fulfil object function, their typical post-verbal position can be determined by dominant SVO order which is characteristic of the majority of European languages within the Standard Average European area (cf. [33]) and in the production of a second language it is preferred by L2 learners regardless of basic word order in their native language (cf. [34, p. 87]).

ACKNOWLEDGEMENTS

This work was supported by the Slovak Research and Development Agency within contract No. APVV-19-0155 Language Errors in Slovak as a Foreign Language Based on Learner Corpus.

[11] Although the ratio of any kind of enclitic error decreases at higher proficiency level, this happens in favour of increasing frequency of another kind of error. With this fact in mind, it would not be quite correct to assume that, for higher proficiency speakers, the placement of enclitic component no longer constitutes a challenge to overcome. A good example are copular errors, the frequency of which declines compared to other enclitic errors at higher acquisition level, but does not decline with respect to the total number of copulas in texts (cf. 8 erroneously placed copulas out of 79 copulas for Slavic A1 – B1 texts, and 7 erroneously placed copulas out of 65 for Slavic B2 – C1 texts; 9 errors out of 67 copulas for Non-Slavic A1 – B1 texts, and 6 errors out of 28 copulas in Non-Slavic B2 – C1 texts).

References

[1] Hana, J. (2007). Czech Clitics in Higher Order Grammar. Columbus: The Ohio State University, 281 p.

[2] Uhlířová, L. (1987). Knížka o slovosledu. Praha: Academia, 160 p.

[3] Ellis, R., and Barkhuizen, G. (2005). Analysing Learner Language. Oxford: Oxford University Press, 416 p.

[4] Mistrík, J. (1966). Slovosled a vetosled v slovenčine. Bratislava: Vydavateľstvo slovenskej akadémie vied, 276 p.

[5] Kačala, J. (2013). Syntagmatický slovosled v slovenčine. Martin: Vydavateľstvo Matice slovenskej, 220 p.

[6] Franks, S., and King, T. H. (2000). A Handbook of Slavic clitics. Oxford: Oxford University Press, 424 p.

[7] Beličová, H., and Uhlířová, L. (1996). Slovanská věta. Praha: Euroslavica, 277 p.

[8] Uhlířová, L., Kosta, P., and Veselovská, L. (2017). Klitikon. In P. Karlík, M. Nekula and J. Pleskalová (eds.), Nový encyklopedický slovník češtiny. Masarykova univerzita, Brno. Accessible at: https://www.czechency.org/slovnik/KLITIKON.

[9] Starý Kořánová, I. (2017). Příklonky a vazaly infinitivu. Studie z aplikované lingvistiky, 11 (special issue), pages 109–120.

[10] Omnia Slovaca III Maior. Accessible at: http://ske.juls.savba.sk/bonito/index.html, (Cited 17.03.2021).

[11] Avgustinova, T., and Oliva, K. (1995). On the nature of the Wackernagel position in Czech. In U. Junghanns and G. Zybatow (eds.), Formale Slavistik, pages 25–47, Vervuert, Frankfurt am Main.

[12] Junghanns, U. (2002). Clitic climbing im Tschechischen. Linguistische Arbeitsberichte, 2002(80), pages 57–90.

[13] Kosek, P. (2011). Enklitika v češtině barokní doby. Brno: Host, 360 p.

[14] Palková, Z. (1994). Fonetika a fonologie češtiny. Praha: Karolinum, 367 p.

[15] Slovenský národný korpus – errkorp-0.1. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2019. (Pre-pilot version not accesible for public.).

[16] Ivanová, M., Kyseľová, M., and Gálisová, A. (2021). Acquiring word order in Slovak as a foreign language: comparison of Slavic and Non-Slavic learners utilizing corpus data [Data set]. Zenodo. Accessible at: http://doi.org/10.5281/zenodo.4625429.

[17] Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), pages 37–46.

[18] Glen, S. "Cohen's Kappa Statistic" From StatisticsHowTo.com: Elementary Statistics for the rest of us! Accessible at: https://www.statisticshowto.com/cohens-kappa-statistic/, (Cited 05.03.2021).

[19] Ortega, L. (2003). Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing. Applied Linguistics, 24(4), pages 492–518.

[20] De Clercq, B., and Housen, A. (2017). A Cross-Linguistic Perspective on Syntactic Complexity in L2 Development: Syntactic Elaboration and Diversity. The Modern Language Journal, 101(2), pages 315–334.

[21] Klein, W., and Perdue, C. (1997). The Basic Variety (or: Couldn't natural languages be much simpler?). Second Language Research, 13(4), pages 301–347.

[22] Byrnes, H. (2009). Emergent L2 German writing ability in a curricular context: A longitudinal study of grammatical metaphor. Linguistics and Education, 20(1), pages 50–66.
[23] Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Kortmann and B. Szmrecsanyi (eds.), Linguistic complexity: Second language acquisition, indigenization, contact, pages 127–155, De Gruyter, Berlin.
[24] Biber, D., Nekrasova, T., and Horn, B. (2011). The Effectiveness of Feedback for L1-English and L2-Writing Development: A Meta-Analysis. ETS Research Report Series, 2011(1), pages i–99.
[25] Biber. D., and Gray, B. (2011). Grammatical change in the noun phrase: The influence of written language use. English Language and Linguistics, 15(2), pages 223–250.
[26] Norris, J. M., and Ortega, L. (2009). Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. Applied Linguistics, 30(4), pages 555–578.
[27] Bardovi-Harlig, K., and Bofman, T. (1989). Attainment of Syntactic and Morphological Accuracy by Advanced Language Learners. Studies in Second Language Acquisition, 11(1), pages 17–34.
[28] Wolfe-Quintero, K., Inagaki, S., and Kim, H.-Y. (1998). Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity. University of Hawai'i, Second Language Teaching and Curriculum Centre, Honolulu.
[29] Vyatkina, N. (2012). The Development of Second Language Writing Complexity in Groups and Individuals: A Longitudinal Learner Corpus Study. Modern Language Journal, 96(4), pages 576–598.
[30] Amzallag, O. (2019). 3rd Person Object Pronoun Use in French Beginning Textbooks. Journal of Second and Multiple Language Acquisition – JSMULA 7(1), pages 15–47.
[31] Viberg, Å. (2002). Basic Verbs in Second Language Acquisition. Revue française de linguistique appliquée, 7(2), pages 61–79.
[32] Manova, S. (2015). Affix Order and the Structure of the Slavic Word. In S. Manova (ed.), Affix Ordering Across Languages and Frameworks, pages 205–230, Oxford University Press, New York.
[33] Haspelmath, M. (2001). The European linguistic area: Standard Average European. In M. Haspelmath (ed.), Language typology and language universals. (Handbücher zur Sprach- und Kommunikationswissenschaft), pages 1492–1510, de Gruyter, Berlin.
[34] Odlin, T. (1997; first published 1989). Language Transfer. Cross-linguistic influence in language learning. Cambridge: Cambridge University Press, 210 p.

sciendo
DOI 10.2478/jazcas-2021-0034

SYSTEMIC AND NON-SYSTEMIC VALENCY BEHAVIOR OF CZECH DEVERBAL ADJECTIVES

VERONIKA KOLÁŘOVÁ – ANNA VERNEROVÁ
– JANA KLÍMOVÁ
Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

KOLÁŘOVÁ, Veronika – VERNEROVÁ, Anna – KLÍMOVÁ, Jana: Systemic and non-systemic valency behavior of Czech deverbal adjectives. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 371 – 382.

Abstract: We present results of an automatic comparison of valency frames of interlinked adjectival and verbal lexical units based on the valency lexicons NomVallex and VALLEX. We distinguish nine derivational types of deverbal adjectives and examine whether they tend to display systemic or non-systemic valency behavior. The non-systemic valency behavior includes changes in the number of valency complementations and, more dominantly, non-systemic forms of actants, especially a prepositional group.

Keywords: deverbal adjective, derivational type, non-systemic valency, passive valency

1 INTRODUCTION

Similarly to valency of verbs and deverbal nouns, valency of deverbal adjectives (DAs) plays an important role in the syntactic structure of a sentence ([1], [2], [3], [4]). It is typical of the surface syntactic structure of deverbal adjectives that one valency slot of their base verb occupies a position of a noun being modified by the given adjective ([5]; Sect. 2.2). Which particular valency slot turns into the governing noun depends on the derivational type of the DA (Sect. 2.1 and 2.2). For example, both *podezírající* 'suspecting' (2) and *podezíraný* 'suspected' (3) are derived from the verb *podezírat* 'to suspect' (1), but the governing noun of the former corresponds to the verbal Actor (ACT; i.e., who is suspecting), and of the latter to the verbal Addressee (ADDR; i.e., who is suspected). The valency complementations that remain in the valency frame (VF) of a DA are expected to inherit morphemic forms from the corresponding valency complementations of their base verb (([6], [7]); cf. the same forms of Patient (PAT) in (1–3) and the same form of ADDR in (1–2)), or to change in a regular way, cf. the predictable change Nom > Ins / *od* 'from'+Gen in (1) and (3).

(1) *podezírat* 'to suspect': ACT(Nom) ADDR(Acc) PAT(*z*+Gen,že)
policie.ACT ***podezírá*** *politika*.ADDR *z podvodu*.PAT / *že podvádí*.PAT
'the police suspects a politician of a fraud / that he is swindling'

(2) *podezírající* ‘suspecting’: ADDR(Acc) PAT(*z*+Gen,že)
policie ***podezírající*** *politika*.ADDR *z podvodu*.PAT */ že podvádí*.PAT
‘police suspecting a politician of a fraud / that he is swindling’

(3) *podezíraný* ‘suspected’: ACT(Ins,*od*+Gen) PAT(*z*+Gen,že)
politik ***podezíraný*** *policií*.ACT *z podvodu*.PAT */ že podvádí*.PAT
‘a politician suspected by the police of a fraud / that he is swindling’

However, various irregularities (described in the literature only cursorily ([6], [8])) are rather common, including changes in the number of valency complementations, cf. the only valency slot in the VF of the DA *podezřelý-1* ‘suspect(ed)’ (4a), as well as non-predictable changes in morphemic forms of some valency complementations, cf. the forms of ACT and PAT of the adjective *podezřelý-2* ‘suspect’ (4b) and the forms of ADDR and PAT of the adjective *podezíravý* ‘suspectful’ (5) with those in (1).

(4) a. *podezřelý-1* ‘suspect(ed)’: PAT(*z*+Gen,že)
politik ***podezřelý*** *z podvodu*.PAT */ že podvádí*.PAT
‘a politician suspect(ed) of a fraud / that he is swindling’

b. *podezřelý-2* ‘suspect’: ACT(Dat) PAT(Ins)
politik ***podezřelý*** *policii*.ACT *potenciálně podvodným jednáním*.PAT
‘a politician suspect to the police due to [lit. by] a potentially fraudulent act’

(5) *podezíravý* ‘suspectful’: ADDR(*k*+Dat) PAT(že)
policie ***podezíravá*** *k politikovi*.ADDR, *že podvádí*.PAT / **z podvodu*.PAT
‘police suspicious towards a politician, that he is swindling / *of a fraud’

In this paper, we exploit manually annotated valency properties of Czech deverbal adjectives and verbs captured in valency lexicons NomVallex [9] and VALLEX [10] (Sect. 2). Drawing on an automatic comparison between VFs of DAs and their base verbs enables us to specify regular (systemic) and irregular (non-systemic) valency behavior of deverbal adjectives (Sect. 3). We provide the first statistical data and show what kinds of non-systemic valency behavior of DAs are the most frequent and what derivational adjectival types are subject to non-systemic changes to the largest extent (Sect. 4).

2 DEVERBAL ADJECTIVES IN THE NomVallex LEXICON

NomVallex is a valency lexicon of Czech deverbal nouns [9], adjectives and deadjectival nouns; it is based on the theoretical framework of the Functional

Generative Description (FGD) and on corpus data (Czech National Corpus, subcorpus SYNv8 [11], and Araneum Bohemicum Maximum [12]). Each lexical meaning of an adjective is treated as one lexical unit (LU) of a lexeme. Applying the valency theory of the FGD [13], valency properties of a LU are captured in a valency frame, which is modeled as a sequence of valency slots, supplemented with their morphemic forms. The following types of complementations may fill in the individual slots of VFs of most deverbal adjectives: obligatory or optional actants, i.e., Actor (ACT), Patient (PAT), Addressee (ADDR), Effect (EFF), and Origin (ORIG), e.g., *chtivý peněz*.PAT 'avid for money', *prodejný mládeži*.ADDR 'marketable to the youth', *odvolatelný z funkce*.ORIG 'dismissible from the post', and obligatory free modifications, especially those with the meaning of direction, e.g., *stěna přilehlá ke kostelu*.DIR3 'a wall adjoining to the church'.

NomVallex currently contains 258 adjectival LUs in 160 lexemes, out of which 195 LUs in 128 lexemes are considered to be deverbal. Deverbal adjectives are classified into nine types (Sect. 2.1, Table 1), and where possible, valency frames of particular LUs are linked to valency frames of their base verbal LUs in the VALLEX lexicon (164 adjectival LUs, 6 of which are linked to more than one verb, Sect. 4).

In NomVallex, all Czech deverbal derivates with adjectival inflection are regarded to be deverbal adjectives, no matter whether they denote an action (e.g., *porota rozhodující o cenách* 'jury deciding the awards'), a property (e.g., *rozhodující okamžik* 'decisive moment') or an object (*můj známý* 'an acquaintance of mine').

2.1 Types of deverbal adjectives

We distinguish nine derivational types of deverbal adjectives, exemplified here by DAs derived from verbs *podezřívat* 'suspect' and *rozpadat se – rozpadnout se* 'disintegrate' (Table 1). The classification is somewhat heterogenous: types (i)–(iv) are characterized by mostly regular derivation from transgressive and participial verbal forms, types (v)–(vii) reflect what the adjectives mostly denote, and types (iv') and (viii) are singled out from the preceding groups because of their specific valency properties.[1]

Because types (i)–(ii) are expected to display systemic valency behavior [7] and types (vii)–(viii) usually have no valency complementations (cf. the adjective-noun *podezřelý* 'the suspect' with an empty VF and *zařízení čtoucí dopravní značky* 'a device reading traffic signs' vs. *čtecí zařízení // *zařízení čtecí dopravní značky* 'reading device // device intended for reading *traffic signs'), these DAs are only rarely covered in NomVallex. Instead, types (iii)–(vi) are focused on (Sect. 4).

[1] We take types (i)–(iv) and (v)–(vii) from [14]. However, the relationship between the derivational suffix and DA's meaning is not always straightforward (e.g.. the adjective *rozpadací* 'disintegrating easily' is not an adjective of purpose).

Adjectival type			*podezřívat*^{impf} – *podezírat*^{impf} – **podezřít*^{pf} ‘to suspect’	*rozpadat se*^{impf} – *rozpadnout se*^{pf} ‘to disintegrate’
Derived from	(i)	present transgressive	*podezřívající* – *podezírající* ‘suspecting’	*rozpadající se* ‘disintegrating’
	(ii)	past transgressive	-	*rozpadnuvší se* ‘having disintegrated’
	(iii)	active participle	*podezřelý* ‘suspect (adjective)’	*rozpadlý* ‘disintegrated’
	(iv)	passive participle (perfective or imperfective)	*podezíraný* – *podezřívaný* ‘suspected’	-
	(iv’)	passive participle of a (typically perfective) reflexive intransitive verb	-	*rozpadnutý* ‘disintegrated’
Expressing	(v)	potential to be affected by an action, with the most productive suffix *-telný*	*podezíratelný* – *podezřívatelný* ‘one who can be suspected’	*rozpadnutelný* ‘that which can be disintegrated’
	(vi)	property resulting from a tendency to repeat an action, formed with various suffixes	*podezíravý* – *podezřívavý* ‘suspectful’	*rozpadavý* ‘(prone to) disintegrating’
	(vii)	purpose, mostly using suffix *-cí*, e.g. *krycí* ‘aimed to cover’	-	*rozpadací* ‘disintegrating easily’
	(viii)	a concretum, usually a person (semantically a noun, formally an adjective)	*podezřelý* ‘suspect (noun)’	-

Tab. 1. Types of DAs

2.2 Adjectival passive valency

According to [15], adjectives usually have one valency slot which is filled with the noun they modify (so-called passive valency; e.g., *rozpadlý plot* ‘disintegrated fence’). In case of DAs, it corresponds to a valency slot of the base verb [5] (e.g., the ADDR in constructions *podezírat politika.*ADDR *z podvodu.*PAT ‘to suspect a politician of a fraud’ > *politik podezíraný z podvodu.*PAT ‘a politician suspected of a fraud’). In the VFs of DAs, we treat passive valency as a form of expression of a valency complementation and mark it by an upward arrow, see (6). The distribution

of passive valency across the annotated adjectival types in NomVallex is given in Table 2.

(6) *podezíraný* 'suspected': ACT(Ins,*od*+Gen) PAT(*z*+Gen,že) ADDR(↑)

Type	Number of LUs			
	Passive valency			Total
	ACT(↑)	ADDR(↑)	PAT(↑)	
(i)	9	-	-	9
(ii)	3	-	-	3
(iii)	25	2	1	28
(iv)	9	4	18	31
(iv')	11	-	-	11
(v)	3	-	17	20
(vi)	71	-	15	86
(vii)	-	-	1	1
(viii)	3	1	2	6
Total	134	7	54	195

Tab. 2. Passive valency of DAs in NomVallex

Only types (i) and (ii) systematically use ACT as their passive valency, the others, though preferring one complementation to be their passive valency, allow for exceptions, often caused by unusual valency behavior of their base verb or by uncertainty about their base verbal LU, especially in case of reflexive and non-reflexive verbal variants. Adjectives representing type (iii) logically strongly prefer ACT to be their passive valency (e.g., *problém vzniklý z čeho* 'a problem arising from sth.'), however, also PAT (*škoda vzniklá komu*.ACT 'harm inflicted upon sb.') or ADDR (*podezřelý z čeho* 'suspect(ed) of') are exceptionally possible. Types (iv) and (v) obviously prefer PAT (e.g., *povolaný* 'conscripted') or ADDR (*podezíratelný* 'one who can be suspected') to be their passive valency. However, there are several cases in which ACT stands in this position, e.g., with the adjective *poddajný* 'docile' derived from the reflexive verb *poddat se* 'yield', rather than from its non-reflexive variant *poddat* 'subdue'. Reflexivity of the base verbal lexical unit also leads to using ACT as passive valency with adjectives representing type (iv') (e.g., *odhodlaný* 'determined', *rozhádaný* 'quarreling (e.g., couple)', *zamilovaný* 'in love'). Type (vi) is a heterogenous group of DAs derived by various suffixes and expressing various properties; in our data, these DAs prefer ACT to be their passive valency (e.g., *vnímavý* 'sensitive'). Types (vii) and (viii), i.e., adjectives of purpose and adjectives denoting a person, are too rare in our data to be able to generalize their typical passive valency.

3 SYSTEMIC AND NON-SYSTEMIC VALENCY BEHAVIOR OF DEVERBAL ADJECTIVES

3.1 Systemic valency behavior

Systemic valency behavior of DAs concerns both deep and surface realization of adjectival valency and it differs depending on the adjectival type.

When determining the deep syntactic structure of DAs, i.e., especially members of their valency frames, the adjectives are expected to inherit all actants that are present in the valency frame of their base verbal lexical unit, though one of them is only expressed as passive valency (Sect. 2.2).

As for the surface expression of actants, all morphemic forms which do not change are regarded to be systemic. These include prepositionless cases Gen, Dat, Acc and Ins, an infinitive (Inf), prepositional groups (PGs), conjunctions and content clauses (CONT), expression *jako* 'as'+Nom (e.g., *proslulý jako* 'famous as'), and expressions containing preposition *za* 'as/for' plus an adjective in Acc (*za*+adj-Acc, e.g., *považované za méněcenné* 'considered to be inferior').

There are two verbal morphemic forms that we consider to be subject to systemic changes, namely Nom (7) and prepositionless Acc (8). Changes Nom > Ins / *od* 'from'+Gen are typical of ACT of adjectives belonging to types (iv) [5] and (v), though the change Nom > *od* 'from'+Gen is rather rare.

(7) *hacker*.Nom ***vydírá*** *podnikatele*.Acc
'a hacker is blackmailing an entrepreneur'

(a) Nom > Ins
podnikatel ***vydíraný*** / ***vydíratelný*** *hackerem*.Ins
'an entrepreneur blackmailed / susceptible to blackmailing by a hacker'

(b) Nom > *od* 'from'+Gen
podnikatel ***vydíraný*** / ***vydíratelný*** *od hackera*
'an entrepreneur blackmailed / susceptible to blackmailing from a hacker'

(8) Acc > Gen
znát *poměry*.Acc > ***znalý*** *poměrů*.Gen
'to know the conditions > knowledgeable about the conditions'

While no adjectival complementation can be expressed by Nom, Acc can be subject to the systemic change Acc > Gen but can also remain unchanged, esp. with DAs of types (i), e.g., *znající něco* 'knowing sth.', and (ii), e.g., *poznavší něco* 'having learnt sth.'. Actants of other adjectival types only exceptionally allow expression by an Acc, e.g., *dlužný někomu vysvětlení*.Acc 'owing sb. an explanation'

and *dvě děti naučené jednu roli*.Acc ‘two children who-have-learnt the same role’. Instead, the other adjectival types prefer the systemic change Acc > Gen, e.g., *znalý poměrů*.Gen ‘knowledgeable about the conditions’ and *chtivý peněz*.Gen ‘avid for money’.

3.2 Non-systemic valency behavior

Non-systemic valency behavior of DAs involves three phenomena:

i. a change in the number of slots in the valency frame of an adjective (e.g., the VF of the adjective *chtivý* ‘avid’, which is derived from *chtít* ‘to want’, only contains PAT, e.g., *chtivý peněz*.PAT ‘avid for money’, losing the original verbal ORIG, cf. **chtivý od někoho*.ORIG ‘*avid from sb.’; Sect. 4.1);

ii. non-systemic forms of actants (e.g., Acc > Dat, *vláda poslouchá prezidenta*. Acc > *vláda poslušná prezidentovi*.Dat ‘the government obeys the president > government obedient to the president’; Sect. 4.2 and 4.3);

iii. a change in the nature of a valency complementation to exclusively nominal in the case of adjectives of type (viii); for example, the adjective-noun *známý* ‘acquaintance, friend’ denotes a person and its VF only contains the nominal complementation Appurtenance (e.g., *můj*.APP *starý známý* ‘an old acquaintance of mine’). However, this is extremely rare in the NomVallex data and is not dealt with in the paper.

4 AN AUTOMATIC COMPARISON OF VERBAL AND ADJECTIVAL VALENCY FRAMES

Our automatic comparison of VFs of adjectives in NomVallex and VFs of their base verbs in VALLEX covers 164 verb–adjective pairs. The automatic procedure captures systemic valency behavior (when the number and type of valency slots, including passive valency, is the same in the adjectival VF as in the corresponding verbal VF and their forms are either the same or correspond to a systemic change, see Sect. 3.1) as well as non-systemic valency behavior; its output is captured in the `valdiff` attribute of the DA.

In this Section, we only focus on differences in the number or forms of actants, leaving out free modifications. Going from the deep to the surface valency structure, we first focus on the difference in the number of actants (Sect. 4.1), then we provide the general statistics on the number of systemic and non-systemic morphemic forms in adjectival VFs (Sect. 4.2), and finally we present the distribution of non-systemic adjectival forms (Sect. 4.3).

4.1 Differences in the number of actants

Any change in the number of actants (i.e., in the deep syntactic structure) between a verb and a DA indicates a change in meaning. Table 3 exemplifies DAs with different number of actants and shows that such changes are rather rare, especially when it comes to adding a new actant (e.g., *zuřivý vůči podvodníkům*. ADDR 'furious at the swindlers') or deleting one but not all actants (e.g., *podezřelý z podvodu*.PAT **policií*.ACT 'suspect(ed) of a fraud *by the police'). Deleting actants may affect all types of DAs, even particular LUs of type (i), e.g., *rozhodující okamžik* 'decisive moment'.

Actant	Passive valency	„Standard" valency				
		Shared	Added	Deleted		Total
				EMPTY VF	Non-EMPTY VF	
ACT	107	41	-	11 *dotčený* 'in question'	6 *podezřelý-1* 'suspect(ed)'	58
ADDR	7	11	1 *zuřivý k*+Dat 'furious at sb.'	4 *prodejný* 'corruptible'	4 *způsobilý* 'eligible'	20
EFF	-	8	1 *příjemný čím* 'pleasant by/ due to'	2 *oprávněný* 'justified'	-	11
ORIG	-	5	2 *zkušený z*+Gen 'experienced [lit. from]'	1 *vědomý* 'willful'	5 *chtivý* 'avid'	13
PAT	50	87	-	21 *rozhodující* 'decisive'	1 *přejícný* 'ungrudging'	109
Total	164	152 72%	4 2%	39 18,5%	16 7,5%	211 100%

Tab. 3. A difference in the number of adjectival actants

4.2 Systemic vs. non-systemic morphemic forms

A general statistics on the number of morphemic forms in verbal and corresponding adjectival VFs is given in Table 4. Looking at the average numbers of all systemic and non-systemic forms, we can see that non-systemic forms account for 30% of the total number of adjectival forms in our data; the highest percentage of non-systemic forms is detected with types (vi), 51% (e.g., *podezíravý k politikovi* 'suspectful towards a politician'), and (iv'), 32% (e.g., *dojatý z vyprávění* 'touched by

[lit. from] the story'). However, even DAs of type (vi) may display systemic valency behavior (e.g., *nápomocný* 'helpful', *pamětlivý* 'mindful'). In line with our expectations, types (i) and (ii) only use systemic morphemic forms. In case of types (vii) and (viii) (with 1 and 6 verb-adjective pairs, respectively), only slots corresponding to passive valency are shared between the base verbs and the derived adjectives.

Adjectival type	Verb -adjective pairs	Base verb's forms	Adjectival forms				
			Systemic		Non-systemic		Total (100%)
				%		%	
(i)	9	23	23	100%	0	0%	23
(ii)	3	5	5	100%	0	0%	5
(iii)	14	31	22	75%	7	25%	29
(iv)	30	84	70	86%	11	14%	81
(iv')	9	21	17	68%	8	32%	25
(v)	18	19	19	76%	6	24%	25
(vi)	74	139	57	49%	58	51%	115
Total	157	322	213	70%	90	30%	303

Tab. 4. The number of systemic and non-systemic adjectival morphemic forms of actants shared between the base verb and a derived DA, excluding the passive valency

4.3 Distribution of non-systemic forms of adjectival actants

Our data shows that a prepositional group is the most frequent non-systemic form, documented esp. with PAT (43 instances, e.g., PG *k* 'to'+Dat in (9)), or ACT (11 instances, e.g., PG *pro* 'for'+Acc in (10)). As for a non-systemic PG as an expression of PAT or ADDR, it often occurs in valency frames of adjectives expressing a property which relates to sb./sth., e.g., *vnímavý* 'sensitive', *podezíravý* 'suspectful', *snášenlivý* 'tolerant', see (9) and (11).

(9) ***vnímat*** *potřeby*.PAT *jiných* > ***vnímavý*** *k potřebám*.PAT *jiných*
'to sense the needs of others > sensitive towards the needs of others'

(10) *nevidomí občané*.ACT *mohou* ***vnímat*** *výstražný pás*.PAT >
'blind people can perceive the safety strip' >
výstražný pás ***vnímatelný*** *nevidomými občany*.ACT / *pro nevidomé občany*.ACT
'a safety strip perceivable by the blind people / for the blind people'

(11) ***podezírat*** *politika*.ADDR > ***podezíravý*** *k politikovi*.ADDR
'to suspect a politician > suspectful towards a politician'

The second most frequent non-systemic form is a prepositionless case, surprisingly Dat for both ACT and PAT, cf. (12–14); Gen and Ins are also possible for PAT and EFF, see (12), (15) and (16).

(12) *policie*.ACT ***podezírá*** *politika*.ADDR, *že podvádí*.PAT >
'the police suspects a politician that he is swindling' >
politik ***podezřelý*** *policii*.ACT *potenciálně podvodným jednáním*.PAT
'a politician suspect to the police by a potentially fraudulent act'

(13) ***účastnit se*** *našeho jednání*.PAT > ***účastný*** *našemu jednání*.PAT
'to take part in our talks > taking part in [lit. to] our talks'

(14) *vláda* ***poslouchá*** *prezidenta*.PAT >
'the government is obeying the president' >
vláda ***poslušná*** *prezidentovi*.PAT
'a government obedient to the president'

(15) ***nemůže*** *promluvit*.PAT > *není* ***mocen*** *slova*.PAT
'he isn't able to speak > incapable of a word'

(16) *spisovatel*.ACT ***zná*** *o městě*.PAT *historku*.EFF >
'a writer knows a story about the town' >
město je spisovateli.ACT ***známé*** *historkou*.EFF
'a town known to the writer through a story'

Actants PAT and EFF may be expressed by a non-systemic infinitive or a content clause, cf. (17–18) for PAT.

(17) *osoba* ***odpovědná*** *jednat*.PAT
'sb. responsible to act'

(18) ***spokojený****, že vše dobře dopadlo*.PAT
'content that everything ended well'

Actant	Morphemic form	Adjectival type					Total
		(iii)	(iv)	(iv')	(v)	(vi)	
ACT	Dat	1	-	-	1	3	5
	PG	-	-	-	4	7	11
ADDR	PG	-	-	-	-	3	3
EFF	Ins	-	-	-	-	2	2
	PG	-	2	-	-	1	3
	Inf	-	1	-	-	-	1
	CONT	-	1	-	-	1	2

Actant	Morphemic form	Adjectival type					Total
		(iii)	(iv)	(iv')	(v)	(vi)	
ORIG	PG	-	-	-	1	-	1
PAT	Gen	-	-	-	-	1	1
	Dat	-	-	-	-	3	3
	Ins	1	-	-	-	-	1
	PG	2	7	2	-	32	43
	Inf	1	-	2	-	2	5
	CONT	2	-	4	-	3	9
Total		7	11	8	6	58	90

Tab. 5. A distribution of non-systemic adjectival forms

5 CONCLUSION

We have presented results of an automatic comparison of valency frames of interlinked adjectival and verbal lexical units based on the valency lexicons NomVallex and VALLEX. Differentiating nine types of Czech deverbal adjectives, we have observed that non-systemic valency behavior of deverbal adjectives is mostly manifested by either a difference in the number of actants or non-systemic forms of actants, out of which the non-systemic forms are more dominant, represented especially by a non-systemic prepositional group. While a difference in the number of actants may affect all types of deverbal adjectives, even those derived from present transgressives but not denoting an action (e.g., *rozhodující okamžik* 'decisive moment'), non-systemic forms of actants are only characteristic of selected adjectival types, most significantly of adjectives derived from verbs not strictly regularly and denoting various properties (e.g., *podezíravý* 'suspectful').

ACKNOWLEDGEMENTS

The research reported in the paper was supported by the Czech Science Foundation under the project GA19-16633S. This work has been using language resources developed, stored and distributed by the LINDAT/CLARIAH-CZ project (LM2018101) of the Ministry of Education, Youth and Sports of the Czech Republic.

References

[1] Panevová, J. (1998). Ještě k teorii valence. Slovo a slovesnost, 59, pages 1–14.
[2] Svozilová, N., Prouzová, H., and Jirsová, A. (2005). Slovník slovesných, substantivních a adjektivních vazeb a spojení. Praha: Academia, 580 p.
[3] Kopřivová, M. (2006). Valence českých adjektiv. Praha: Nakladatelství Lidové noviny, 125 p.
[4] Skwarska, K. (2018). Valence adjektiv v komparativním pohledu (na materiálu češtiny,

ruštiny a polštiny). Slavia, 87(1–3), pages 302–315.
[5] Piťha, P. (1982). K otázce valence u adjektiv. Slovo a slovesnost, 43, pages 113–118.
[6] Daneš., F., Hlavsa, Z., Grepl. M. et al. (1987). Mluvnice češtiny 3. Skladba. Praha: Academia, 748 p.
[7] Doležalová, D. (2005). Automatic Construction of a Valency Lexicon of Czech Adjectives. In V. Matoušek, P. Mautner and T. Pavelka (eds.), Text, Speech and Dialogue. TSD 2005. Lecture Notes in Computer Science, vol. 3658, pages 56–60, Berlin/Heidelberg: Springer.
[8] Najbrtová, K. (2017). Korpusová analýza přejímání valenčních rámců u adjektiv derivovaných sufixem -telný. Ph.D. thesis. Brno: Masarykova univerzita, 226 p.
[9] Kolářová, V., Vernerová, A., and Klímová, J. (2020). NomVallex I. Valenční slovník substantiv. Praha: Ústav formální a aplikované lingvistiky, 231 p. Accessible at: http://hdl.handle.net/11234/1-3420.
[10] Lopatková et al. (2020). VALLEX 4.0, LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Prague. Accessible at: http://hdl.handle.net/11234/1-3524.
[11] Křen, M. et al. (2019). Korpus SYN, verze 8 z 12. 12. 2019. Praha: Ústav Českého národního korpusu FF UK. Accessible at: http://www.korpus.cz.
[12] Benko, V. (2015). Araneum Bohemicum Maximum, verze 15.04. Praha: Ústav Českého národního korpusu FF UK. Accessible at: http://www.korpus.cz.
[13] Panevová, J. (1980). Formy a funkce ve stavbě české věty. Praha: Academia, 222 p.
[14] Rusínová, Z. (2016). Deverbální adjektivum. In P. Karlík, M. Nekula and J. Pleskalová (eds.), Nový encyklopedický slovník češtiny, page 328, Praha: Nakladatelství Lidové noviny.
[15] Boguslavsky, I. (2003). On the Passive and Discontinuous Valency Slots. In Proceedings of the 1st International Conference on Meaning-Text Theory, pages 129–138, Paris, Ecole Normale Supérieure.

sciendo
DOI 10.2478/jazcas-2021-0035

TOWARDS CLASSIFICATION OF STATIVE VERBS IN VIEW OF CORPUS DATA

SVETLOZARA LESEVA – IVELINA STOYANOVA
– HRISTINA KUKOVA
Institute for Bulgarian Language Prof. Lyubomir Andreychin, Bulgarian Academy of Sciences, Sofia, Bulgaria

LESEVA, Svetlozara – STOYANOVA, Ivelina – KUKOVA, Hristina: Towards classification of stative verbs in view of corpus data. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 383 – 393.

Abstract: The paper presents work in progress on the compilation and automatic annotation of a dataset comprising examples of stative verbs in parallel Bulgarian-Russian corpora with the goal of facilitating the elaboration of a classification of stative verbs in the two languages based on their lexical and semantic properties. We extract stative verbs from the Bulgarian and the Russian WordNets with their assigned conceptual information (frames) from FrameNet. We then assign the set of probable Bulgarian and Russian stative verbs to the verb instances in a parallel Bulgarian-Russian corpus using WordNet correspondences to filter out unlikely stative candidates. Further, manual inspection will ensure high quality of the resource and its application for the purposes of semantic analysis.
Keywords: stative verbs, parallel corpora, semantic annotation

1 MOTIVATION AND TASK OVERVIEW

Despite the advances in the creation of ever larger corpora, parallel or comparable corpora for specific pairs of languages may still be scarce, especially ones with task-specific labelling. In this paper, we describe a methodology for compiling and annotating a parallel corpus for two Slavic languages, Bulgarian and Russian, tailored to a specific linguistic task: contrastive description of stative verbs.

1.1 Predefined vs. resource-driven classification of stative verbs

Vendler's aspectual classification of verbs into activities, states, achievements and accomplishments [1] subsequently developed and elaborated by Dowty [2] and Van Valin and LaPolla [3], among many others, has provided deep insights into the aspectual nature of situations and predicates. There have been other proposals for classifications according to semantic classes that usually take into account the aspectual class: two such accounts are Paducheva's [4] and Van Valin and Lapolla's [3] classifications. A cursory look at the representation of stative verbs (which we deal with) shows that at certain points (such as predicates of emotion, cognition,

desire), the classifications show substantial similarities, while at others they show different levels of granularity or employ different classes altogether. Table 1 provides a juxtaposition between several very similarly treated classes. We use the original examples in the relevant works.

Paducheva (1996)	Van Valin and LaPolla (1997)
Intention and will: желать 'wish' жаждать 'crave', надеяться 'hope', стремиться *'strive, aspire'*	**Desire**: *want*, *wish*
Temporary emotional states: беспокоиться 'worry', веселиться 'rejoice', возмущаться 'be indignant'	**Internal experience**: experience, feel **Emotions**: *love*, *hate*
Permanent emotional states and relations: любить 'love', обожать 'adore', страдать 'suffer, hurt'	
Mental states: интересоваться 'be interested in', колебаться 'hesitate', знать 'know', помнить 'remember', считать 'consider'	**Cognition**: *know*, *believe*, *understand*
	Propositional attitude: *consider*
	Perception: *hear*, *see*

Tab. 1. Emotion, perception, desire and cognition verbs as presented in [3] and [4]

The classes exemplified represent universally acknowledged semantic distinctions, which nonetheless yield different accounts. A worthwhile effort, which constitutes part of our work, would be to compare relevant classifications with the goal of accommodating meaningful distinctions and enriching the description.

1.2 Stative verbs in language resources

A number of lexical semantic resources, such as FrameNet, WordNet, VerbNet, etc., have employed semantic groupings of different granularity to identify semantically coherent verbs. In WordNet such groupings are defined by semantic primitives that divide the verbal domain into 15 classes [5], while FrameNet provides a much more fine-grained approach based on the definition of conceptual frames [6], cf. Section 2. As a result, distinct semantic classes emerge; as both resources have a netlike structure implemented through a number of relations between the basic units of the resource (synonym sets in WordNet; frames in FrameNet), the so-induced classifications have a partially hierarchical organisation.

Such lexical resources provide a schema for annotating verbs in corpora. At the current stage, we adopt an approach to providing the corpus with as much classificatory information as possible, combining information from different

resources. In order to annotate stative verbs in particular, we need to identify them in the resources. As we employ an alignment between WordNet synsets and FrameNet frames, it is sufficient to identify the relevant verb synsets in WordNet as they are explicitly marked or deduced from the WordNet structure.

1.3 Motivation

Our interest in stative verbs is motivated by a joint project undertaken by Bulgarian and Russian researchers that aims at an ontological description of stative verbs in the two languages. Stative verbs are a natural place to start as in a number of theoretical accounts ([1], [3]) among others, they form one of the building blocks (together with activities) employed in the construal of more complex situations. The features of stative verbs are, nonetheless, far from being exhaustively and definitively determined and the membership of verbs to this ontological class is still subject to debate. The corpus aims at providing a test setting for linguistic observations on stative verbs but may readily be extended to include other ontological classes and to perform other tasks.

2 STATIVE VERBS: DATA SELECTION AND ANNOTATION

2.1 Lexical-semantic resources: WordNet

We combine information from several previously developed resources for Bulgarian and Russian, as well as for English for those resources where the two Slavic languages are linked through it (WordNet).

The Princeton WordNet, PWN [7] is a large-scale lexical database that encompasses the lexis of English organised as a network of synonym sets (synsets) comprising conceptual synonyms (individual members of a synset are called literals) linked to each other by means of conceptual, lexical, and other relations. The semantic description pertaining to a synonym includes a semantic label assigned to each verb or noun synset that denotes the semantic primitive of the respective verb or noun synset [5]. In addition to PWN, we use available wordnets for two other languages, Bulgarian [8] and Russian [9], each of which is mapped onto PWN 3.0 through unique synset identifiers. The Bulgarian WordNet contains 14,103 verb synsets, while the Russian WordNet is considerably smaller with 7,634 verb synsets. These wordnets provide the verb inventories used in the study; corresponding synsets are paired at the synset level through their mapping onto PWN.

Example 1.

ID: eng-30-02756359-v
PWN Synset: {belong:5}
Semantic primitive: verb.stative

Gloss: be a member, adherent, inhabitant, etc. (of a group, organization, or place)
Example: They belong to the same political party.
Bulgarian Synset: {принадлежа:7}
Russian Synset: {принадлежать 02365119}

According to their semantic primitive (or atomic predicate) [5], the verb synsets in WordNet are organised in 15 classes such as verbs of change, verbs of motion, verbs of cognition, verbs of communication, verbs of emotion, among others, and are accordingly labelled at the synset level. The class of stative verbs (marked as verb.stative) includes not semantically coherent verbs, but rather verbs that aspectually belong to the category of states. In addition, stative verbs are also found across other classes, although not necessarily characterised as such (e.g., cognitive, emotion verbs, verbs denoting bodily states, verbs of possession, etc.).

To obtain the verbs that denote states, we assume that the verbs labelled as stative qualify as viable candidates. In the first step, we take the set of stative verb synsets in the Bulgarian WordNet, thus obtaining a collection of 559 synsets. We then expand this number by adding verbs that are hyponyms of stative synsets (659 synsets). Further, we add a selection of verbs labelled with the primitives verb. emotion, verb.cognition and verb.perception, increasing the overall number of synsets to 1,786.

We then match these synsets to their Russian WordNet equivalents whenever they exist. Table 2 shows the number of verb synsets under analysis (and the number of literals they contain) in the Bulgarian and Russian WordNets and their corresponding semantic primitive.

	Bulgarian WordNet		Russian WordNet	
Semantic prime	# synsets	# literals	# synsets	# literals
verb.stative	559	1725	392	641
verb.cognition	503	1776	389	605
verb.perception	342	1173	261	432
verb.emotion	264	1035	210	385
verb.change	27	29	6	10
verb.body	19	86	14	21
Others	72	252	56	96

Tab. 2. Distribution of stative verbs across primes in the Bulgarian and the Russian WordNet

As mentioned above, we also collect verbs whose hypernym is a stative verb but they themselves are assigned a different semantic class. Such synsets can be considered as stative verbs with additional semantic characteristics expressed by the

semantic class assigned: for instance, Example 2 shows two hyponyms of a verb. stative synset where one is classified as verb.consumption (denoting the meaning of a state reached through consumption), and the other, an emotional state, is defined as verb.emotion.

Example 2.

Hypernym ID: eng-30-02604760-v
PWN Synset: {be:4}
Semantic primitive: verb.stative
Gloss: have the quality of being; (copula, used with an adjective or a predicate noun)

Hyponym 1 ID: eng-30-01188342-v
PWN Synset: {be full:1}
Semantic primitive: verb.consumption
Gloss: be sated, have enough to eat

Hyponym 2 ID: eng-30-02604760-v
PWN Synset: {seethe:3; boil:4}
Semantic primitive: verb.emotion
Gloss: be in an agitated emotional state

In the next stages of the research, we intend to focus on ways of increasing (through creating, translating, etc.) the Slavic data with more synsets denoting stative meanings that are frequently found in Bulgarian and Russian parallel, comparable or monolingual corpora, including prefixed verbs that are typical for Slavic languages, but are not included in the Princeton WordNet.

2.2 Lexical-semantic resources: FrameNet

FrameNet [6] is a network of conceptual frames, where each frame represents a script-like description of the conceptual structure of situations, objects or events by means of their participants and props, called Frame Elements [10]. The frames are instantiated by word-meaning pairings called Lexical Units. In addition, frames are linked to each other by means of several hierarchical (Inheritance, Using, Subframe, Perspective) and non-hierarchical relations (Causation, Inchoation, Precedence).

FrameNet frames are assigned to synsets in WordNet using one of the proposed automatic mappings between the resources where lexical units in FrameNet and synonyms in WordNet synsets are aligned; where such alignment is impossible, the synsets are assigned a frame from their parent synset or another suitable frame is assigned using a number of additional automatic procedures [11]. More than 5,000 frame-to-synsets assignments have been validated manually. As conceptual information

is to a large extent language-independent, the semantic information is transferrable across languages. Fig. 1 shows an excerpt of corresponding Bulgarian and Russian synsets labelled as stative either on the basis of their primitive (verb.stative) or as hyponyms of a verb.stative synset. The FrameNet frame assigned to the synsets and the pertaining Frame Elements are also exemplified.

SYNSET	SEM. CLASS	BG SYNSET	RU SYNSET	FRAME	FRAME Elements
eng-30-00033599-v	verb.body	изглеждам:1; имам вид:1; излизам:15	[смотреть, глядеть]	Give_impression	Phenomenon:; Characterization:; Appraisal:; Inference:;
eng-30-00047610-v	verb.body	нося:33	[быть_одетым]	Have_associated	Entity:; Topical_entity:;
eng-30-00047745-v	verb.body	имам:25; нося:34	[носить]	Wearing	Wearer:Sentient; Clothing:Artifact; Body_part:Body_part;
eng-30-00065370-v	verb.body	боледувам:1; страдам:2; болен съм:1; имам:11	[обладать, иметь]	Medical_conditions	Patient:Living_thing; Ailment:;
eng-30-00077698-v	verb.body	задавям се:3; задавя се:3; давя се:3; задушавам се:1	[задыхаться]	Perception_body	Experiencer:Sentient; Body_part:Body_part;
eng-30-00623006-v	verb.cognition	поставям:4; поставя:4; задавам:1; задам:1	[говорить_загадками]	Stimulate_emotion	Experiencer:Sentient; Stimulus:;
eng-30-01188144-v	verb.consumption	умирам от глад:2; умра от глад:2; гладувам:3	[голодать, быть_голодным]	Perception_body	Experiencer:Sentient; Body_part:Body_part;
eng-30-01763101-v	verb.emotion	преливам:3; прелея:3; преизпълнен съм:1; кипя:5	[переливаться_через _край]	Emotion_heat	Experiencer:Sentient; Emotion:; Seat_of_emotion:;
eng-30-01763303-v	verb.emotion	вря:4; бушувам:6	[вызывать_брожение]	Emotion_heat	Experiencer:Sentient; Emotion:; Seat_of_emotion:;
eng-30-02134927-v	verb.perception	звуча:1; прозвучавам:1; прозвуча:1; озвучавам:1	[издавать_звук, звучать]	Give_impression	Phenomenon:; Characterization:; Appraisal:; Inference:;
eng-30-02603699-v	verb.stative	съществувам:3; съм:13; има:4; битувам:1	[быть, существовать]	Existence	Entity:;
eng-30-02614181-v	verb.stative	живея:2; жив съм:1; съм:18	[существовать]	Dead_or_alive	Protagonist:Sentient; Figure:;
eng-30-02614387-v	verb.stative	живея:8; поминувам:1; поминавам:1; помина:1	[жить]	Manner_of_life	Experiencer:; Lifestyle:; Manner:;
eng-30-02198234-v	verb.perception	струва ми се:2; сторва ми се:2; стори ми се:2	[казаться, представляться]	Unattributed_information	Unattributed_information 1 Reported_fact:;

Fig. 1. A sample of the synset-to-frame alignment for several stative verbs

In order to expand the number of stative verbs, we use both the manually checked WordNet-to-FrameNet alignment and the netlike frame organisation, in particular part of the FrameNet frame-to-frame relations. The FrameNet 'tree' stemming from the frame State represents stative situations according to the relation of Inheritance.

Our working assumption is that frames inheriting from State must also be stative, as well as the verbs they describe (Fig. 2). We thus consider a set of 178 frames regarded as describing stative verbs and situations. Some frames can cover both stative and active verbs, e.g., Assessing, which is assigned to verbs such as *value* (stative) and *grade* (active).

Through their alignment with English, the synsets in the Bulgarian and the Russian WordNets are also assigned a frame, confer Table 3 for the most frequent frames. The last column signifies whether the frame is a confirmed stative frame, marked with an X (either coming from the FrameNet tree rooted in State, or manually confirmed as stative).

- State, ID# 150. Definition. (Total related: 134)
  - 132 frames related to State via Inheritance (20 immediate Inheritance, 54 total Inheritance).
    - Locative_relation, ID# 199. Definition. (contains 16 Inheritance relations)
    - Process_initial_state, ID# 152. Definition. (contains 2 Inheritance relations)
    - Change_of_state_initial_state, ID# 185. Definition
    - Change_of_state_endstate, ID# 186. Definition
    - Dead_or_alive, ID# 310. Definition. (contains 1 Inheritance relation)
    - Bearing_arms, ID# 470. Definition.
    - Being_located, ID# 960. Definition. (contains 3 Inheritance relations)
    - State_of_entity, ID# 1600. Definition. (contains 1 Inheritance relation)
    - Being_in_operation, ID# 880. Definition
    - Attention, ID# 791. Definition.
    - Existence, ID# 660. Definition. (contains 1 Inheritance relation)
    - Posture, ID# 18. Definition
    - Being_attached, ID# 307. Definition.
    - Emotions, ID# 1712. Definition. (contains 8 Inheritance relations)
    - Process_completed_state, ID# 234. Definition. (contains 1 Inheritance relation)
    - Process_uncompleted_state, ID# 1764. Definition. (contains 1 Inheritance relation)
    - Thriving, ID# 1771. Definition
    - Dying, ID# 2055. Definition.
    - Transportation_status, ID# 2645. Definition
    - Chaos, ID# 2705. Definition

Fig. 2. The shallow hierarchy beginning with State according to the Inheritance relation

FrameNet frame assigned	# synsets BG	# synsets RU	Stative frame
NO FRAME ASSIGNED	154	92	
Stimulate_emotion	132	111	X
Locative_relation	37	27	X
Categorization	38	32	
Assessing	24	11	
Purpose	22	16	
Perception_body	20	13	
Compatibility	16	11	X
Similarity	15	11	X
Have_associated	15	8	X
Give_impression	11	5	X
Posture	11	9	X
Residence	11	9	X
Existence	10	6	X
Expertise	10	6	X

Tab. 3. The most frequent frames (10+ examples) assigned to stative verbs in the two wordnets

The semantic information from WordNet and FrameNet, including the gloss, semantic prime, examples, etc. from the WordNet synsets and the frame definition, lexical units (other verbs) assigned to the frame, etc. provide valuable semantic information that will be used in the analysis of the stative verbs. The synset semantic primes and the FrameNet frames (as distinct entities from other frames) are especially helpful as they suggest meaningful classificatory categories. For instance, Being_ located (*sit, lie, stand 'be located at'*, etc.) and Spatial_contact (*meet*, *contact*, *touch, adjoin 'be in physical contact with'*) verbs may be defined as subcategories of a more general classification category Location; Residence verbs (*live*, *occupy*, *dwell*, *camp*, *bivouac*, *room*, *stay*, *squat*, *lodge*) and Existence verbs (*live*, *exist*, *be 'have existence'*; *consist in*, *lie in*, *dwell* 'originate in') may be defined as distinct categories, etc. In addition, the analysis of the Frame Elements (the last column of Fig. 1) is very helpful in identifying the semantic and selectional properties of the verbs' arguments.

2.3 Corpus data and preliminary annotatiton

For the purposes of the current work, we employed the Polish-Bulgarian-Russian Corpus ([12], [13]), a parallel corpus for the three languages incorporated in the CLARIN framework. The Corpus consists of 55 parallel texts, comprising 2.23 mln. words for Bulgarian and 2.04 mln. words for Russian from several text genres such as fiction, instruction manuals and technical documentation, legal texts, etc. The parallel texts are automatically aligned at sentence-level and the annotations have been post-edited manually. For the two languages under study we thus obtain 89,562 parallel sentences.

The Bulgarian corpus was POS-tagged using the Bulgarian Language Processing Chain [14]. The Russian part of the corpus was POS-tagged with an available UDPipe language model for Russian [15]. The tagging is necessary in order to identify the relevant verb lemmas to the end of matching them to possible WordNet senses.

3 DATASET OF ANNOTATED EXAMPLES OF STATIVE VERBS

The task is to annotate the stative verbs in the parallel Bulgarian-Russian Corpus obtained from the Polish-Bulgarian-Russian Corpus. The annotation involves the assignment of a relevant WordNet synset that best describes the sense using the Bulgarian and the Russian WordNets. As sense annotation is very sensitive and prone to mistakes, the decision making will be delegated to human experts who will choose the most relevant sense (synset) out of a number of automatically assigned synsets. To facilitate the process, we have adopted a procedure for filtering out non-relevant synsets, which we describe below.

Step 1. We first assign all the possible senses to the lemmatised verbs in the Bulgarian part of the parallel corpus that have at least one stative sense in WordNet. These verbs (or rather their graphic form) have counterparts in the collection of possible relevant synsets and are thus potentially stative.

Step 2. For each sense assigned to a potential stative verb in the Bulgarian part of the corpus, we collect the corresponding synsets from the Russian WordNet, where available.

Step 3. We identify the verbs in the tagged Russian part of the corpus that may potentially belong to the same synset as the corresponding Bulgarian verb in the parallel Bulgarian sentence. The task boils down to finding the intersection of the set of Bulgarian and Russian synsets which are assigned to a Bulgarian verb and a Russian verb, respectively, found in a pair of equivalent sentences: candidates from corresponding synsets appearing in a pair of parallel sentences are very likely translational equivalents.

Step 4. If no pair of verbs from the corresponding Bulgarian and Russian synsets are identified, for each Bulgarian verb, we extract all Russian stative verb translations (in the corresponding sentence) and include them in the list of possible candidates. The assumption is that a state is more likely to be expressed by stative verbs in both languages, even if not from the same synset.
At this stage, a number of heuristics based on semantic relations between synsets can be employed in order to improve filtering of invalid suggestions and reduce further manual validation.

Step 5. After the list of possible senses is reduced through the filtering procedures, we assign the FrameNet frames mapped to the relevant synsets.
As a result, the potentially stative verbs in the Bulgarian-Russian parallel corpus are assigned a number of (filtered-out) senses. Each verb is thus supplied with semantic information derived from the respective WordNet synsets and the assigned FrameNet frames: the semantic prime and the description of the conceptual frame as well as the semantic relations with other synsets or frames. The corpus is then ready to be further disambiguated by human experts.
Initially, we extracted over 30,000 pairs of parallel sentences from the corpus, which were then filtered down to 7,568 examples representing possible stative verbs in Bulgarian and their parallel equivalents in Russian (Example 3).

Example 3.

BG verb: съвпадам
BG sentence: – Вашият разказ е изключително интересен, професоре, въпреки че далеч не ***съвпада*** *с евангелските.*

EN translation: 'Your story is extremely interesting, Professor, though it does not ***coincide*** *at all with the Gospel stories.'*

Potential synsets*:*

eng-30-02658734-v verb.stative {съвпадам:4; съвпадна:4} {coincide:2} 'be the same'

Frame: Compatibility; FEs: Item_1; Item_2; Items; Parameter

eng-30-00345312-v verb.change {съвпадам:1; съвпадна:1} {concur:1; coincide:1} 'happen simultaneously'

No frame assigned

eng-30-02660442-v verb.stative {съвпадам:3; съвпадна:3} {coincide:3; co-occur:1; cooccur:1} 'go with, fall together'

Frame: Existence; FEs: Entity

RU verb: совпадать

RU sentence: – Ваш рассказ чрезвычайно интересен, профессор, хотя он и совершенно не совпадает с евангельскими рассказами.

Potential synsets*:*

eng-30-02658734-v verb.stative {совпадать 02278040}

Frame: Compatibility Item_1:; Item_2:; Items:; Parameter:;

eng-30-00345312-v verb.stative {совпадать 00297090}

No frame assigned

eng-30-02660442-v verb.stative {совпадать 02279659}

Frame: Existence; FEs: Entity

4 CONCLUSIONS

The research presented in this paper suggests several lines of improvement: (i) expanding the inventory of verbs, the FrameNet-to-WordNet alignment, and the size of the parallel corpus; (ii) perfecting the automatic sense assignment and filtering procedures; (iii) outlining major classification categories on the basis of analysis informed both from theoretical work on verb classification and the semantic knowledge encoded in lexical-semantic resources. Further, the classification scheme can be applied to (semi)automatic classification of corpus examples and can be used as a starting point towards automatic semantic role labelling and word sense disambiguation.

ACKNOWLEDGEMENTS

This research is carried out as part of the project *An Ontology of Stative Situations in the Models of Language: a Contrastive Analysis of Bulgarian and Russian* funded by the Bulgarian National Science Fund under the Programme for Bilateral Cooperation, Bulgaria – Russia 2019 – 2020, Grant Agreement No. КП-06-ПРУСИЯ-78 from 2020.

References

[1] Vendler, Z. (1957). Verbs and times. The Philosophical Review, 66(2), pages 143–160.

[2] Dowty, D. (1979). Word meaning and Montague grammar. Reidel: Dordrecht.

[3] Van Valin, R. D., and LaPolla, R. J. (1997). Syntax: structure, meaning, and function. Cambridge: Cambridge University Press.

[4] Paducheva, E. V. (1996). Semanticheskie issledovaniya. Semantika vremeni i vida v russkom yazyke. Semantika narrativa, 464 p.

[5] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1991). Introduction to WordNet: an on-line lexical database. International Journal of Lexicography 3(4), pages 235–244. Revised: 1993. Accessible at: https://wordnetcode.princeton.edu/5papers.pdf.

[6] Baker, C. F., and Ruppenhofer, J. (2002). FrameNet's frames vs. Levin's verb classes. In Proceedings of the 28th annual meeting of the Berkeley.

[7] Fellbaum, C. (Ed.) (1998). WordNet: an electronic lexical database. Cambridge, MA: MIT Press.

[8] Koeva, S. (2010). Bulgarian WordNet – current state, applications and prospects. In Bulgarian-American Dialogues, pages 120–132, Sofia: Prof. M. Drinov Academic Publishing House.

[9] Gelfenbeyn I., Goncharuk A., Lehelt V., Lipatov A., and Shilo V. (2003). Automatic translation of WordNet semantic network to Russian language. In Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2003. Accessible at: http://wordnet.ru.

[10] Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., Baker, C. F., and Scheffczyk, J. (2016). FrameNet II: extended theory and practice. Accessible at: https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf.

[11] Leseva, S., and Stoyanova, I. (2020). Beyond lexical and semantic resources: linking WordNet with FrameNet and enhancing synsets with conceptual frames. In Koeva, S. (Ed.) Towards a semantic network enriched with a variety of semantic relations. Sofia: Prof. Marin Drinov Academic Publishing House of Bulgarian Academy of Sciences.

[12] Kisiel, A., Koseska-Toszewa, V., and Kotsyba, N. (2016). Polish-Bulgarian-Russian parallel corpus, CLARIN-PL digital repository. Accessible at: http://hdl.handle.net/11321/308.

[13] Sosnowski, W. (2016). The parallel Polish-Bulgarian-Russian corpus: problems and solution. In Pastor, G. C. (Ed.), Computerised and corpus-based approaches to phraseology: monolingual and multilingual perspectives, pages 339–349, Geneva: Tradulex. Accessible at: https://www.researchgate.net/publication/319243154_The_parallel_Polish-Bulgarian-Russian_corpus_problems_and_solution.

[14] Koeva, S., and Genov, A. (2011). Bulgarian language processing chain. In Proceeding of the Workshop Integration of multilingual resources and tools in Web applications, Hamburg.

[15] Straka, M., and Straková, J. (2019). Universal dependencies 2.5 models for UDPipe (2019-12-06), LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Accessible at: http://hdl.handle.net/11234/1-3131.

sciendo
DOI 10.2478/jazcas-2021-0036

USAGE AND EMPIRICAL PRODUCTIVITY OF INTERNATIONAL ADJECTIVAL SUFFIXES IN SLOVAK BASED ON GENERAL AND SPECIALISED CORPORA

JANA LEVICKÁ
Slovak National Corpus, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia

LEVICKÁ, Jana: Usage and empirical productivity of international adjectival suffixes in Slovak based on general and specialised corpora. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 394 – 404.

Abstract: The paper attempts to identify the usage and productivity of five different international suffixes in Slovak by means of corpus evidence. The analysis focuses on real and potential productivity in a two-stage comparison: 1) tokens/lemmas occurring in a general balanced corpus vs general corpus of specialised and academic texts, 2) general corpus of specialised and academic texts vs specialised (sub)corpora of medical, legal, economic and religious texts. The aim of the analysis is to explore whether productivity varies across registers by means of statistical measures.

Keywords: productivity, realized productivity, potential productivity, general corpus, specialised corpus, adjective, suffix

1 INTRODUCTION

In the last two decades, terminology research has seen the emergence of a new research topic: variation analysis. One of the most variant-productive areas is that resulting from the clash of two contending tendencies: internationalisation and naturalisation, i.e., coining of new terms from national language resources. In Slovak, usage of the so-called international loanwords[1] has deep historical roots, as Latin was the official language of the Hungarian Kingdom, territory which until 1867 also included Slovakia. A high proportion of words from Latin can be found also in the general Slovak lexicon [1, p. 81], therefore, many Slovaks find borrowing and usage of international words and terms (especially via English) natural.

However, there has always been a natural tendency to coin Slovak counterparts to international words, the realm of terminology including. The clash between these internationalising and naturalising tendencies often results in competing or coexisting (synonymous) words and terms [2, p. 273]. It has been observed that this nationalising tendency is not uniformly present in specialised domains [3, p. 174]; its influence

[1] Terms of Latin and/or Greek origin, occurring at least in three genetically unrelated languages, which are more or less adapted to Slovak [4, p. 89]. More information on corpora can be found in part 2.

manifests itself unevenly, depending on the tradition and character of individual disciplines or domains, as well as on their linguistic and word-forming specificities.

Furthermore, in addition to the traditional preoccupation with nouns, terminology research has also shifted its attention to other parts of speech. It was during the 1990s and after 2000 that adjectives began to emerge into the limelight. Their analyses, underlying differentiating and classifying functions, as well as their ability to *build* multi-word terms, started to be published. The significance of adjectives in terminology and language for specialised purposes (LSP) can be supported also by corpus evidence. While Czech data indicate the presence of as much as 13% of adjectives in LSP texts of SYN 2000 compared to other parts of speech [5], the ratio of adjectives in the majority of specialised corpora of the Slovak National Corpus project amounts only to no more than 9%. However, as can be seen in the table 1, this ratio is at least 2% higher when compared to the reference corpus. Only religious corpus blf-2.0 features roughly the same percentage distribution of adjectives as the reference corpus (7.46%). The third row of table 1 presents the frequency and ratio of gerunds that are also used with differentiating and classifying functions in multi-word terms.

corpus	prim-7.0-frk	prim-9.0-public-prf	prim-9.0-juls-all (MED)	legal-1.1.ver.bz 1991_2011	blf-2.0	ecn-2.0-public
tokens/ corpus	253,137,609	149,581,785	7,099,555	33,600,183	65,920,357	164,987,015
tokens/ adjectives	19,090,396 7.54%	13,415,554 8.97%	660,025 9.3%	4,841,400 9.88%	4,914,860 7.46%	15,540,381 9.42%
tokens/ gerunds	3,174,567 1.25%	2,394,914 1.6%	112,164 1.58 %	1,267,598 2.59%	761,719 1.16%	2,154,124 1.31%

Tab. 1. Number of tokens and ratios of adjectives and gerunds in reference and specialised corpora of the SNC project

Coexisting and competing forms can also be found among adjectives in multi-word terms or in specialised discourse as such. In general, these forms include Slovak counterparts to international adjectives (e.g. *vnútrožilový – intravenózny* 'intravenous'), the latter group comprises also a subgroup that shows the same two tendencies by means of variation of international and Slovak affixes combined with the same international roots (e.g., *bakteri**álny** – baktéri**ový*** 'bacterial').

2 RESEARCH AIMS AND DATA

This paper is an attempt to identify the usage and productivity of five different international suffixes in Slovak by means of corpus evidence. Moreover, the analysis

will focus on a two-stage comparison: 1) tokens/lemmas occurring in a general balanced corpus vs general corpus of specialised and academic texts, 2) general corpus of specialised and academic texts vs specialised (sub)corpora of medical, legal, economic, and religious texts. These corpora may help to explore whether productivity varies across registers.

Suffixes selected for analysis are representative of a minor group within a range of adjectival suffixes.[2] In terms of their composition, the five suffixes may be termed 'reduplicated' as they consist of an (adapted) international adjectival suffix combined with semantically equivalent Slovak **-ný/ny**: **-ál**ny, **-ár**ny, **-i**tný, **-ív**ny, **-óz**ny. Obviously, all of them come from Latin suffixes used to coin adjectives from Latin nouns (**-alis** with the variant **-aris**, -ītus, -ōsus) or verbs (-ōrius). It is worth pointing out, however, that many Slovak adjectives with analysed suffixes entered the Slovak lexicon via French and English.

2.1 Corpora used in the analysis

All three corpora and three subcorpora used in this analysis were released by the Department of the SNC in 2013–2020 and are accessible for all registered users.

The first one, the reference corpus prim-7.0-frk [7], amounting to more than 253 million tokens, is composed of an even share of journalistic, specialised, and fictional texts (64.12% of them are Slovak while 29.51% represent translations) written from 1991–2015. The corpus was used in the compilation of two frequency dictionaries of Slovak (2017, 2018), as well as the reverse dictionary (2018).

The second corpus, prim-9.0-public-prf [8], is a publicly available subcorpus of the primary corpus of the SNC project. Compiled from specialised, academic, and non-fiction texts, this subcorpus features more than 149 million tokens, and documents general discourse of science and research, including specialised journalism. Its texts were written between 1955 and 2019.

The smallest subcorpus of all searched corpora exists as a result of filtering the primary corpus of the SNC project [9], version 9.0. It consists of texts that belong to the field of medicine, written from 1976 to 2019, and comprises slightly more than 7 million tokens.

In order to create a comparable source with other specialised corpora and the reference corpus, the specialised corpus legal-1.1 [10], built in cooperation with the Slovak Ministry of Justice, was narrowed down to legislative texts created between 1991 and 2011. Its approximately 49 million tokens were thus reduced to 33.5 million tokens.

The specialised corpus – blf-2.0 [11] – focusing on the field of religion – was released in 2014. Its texts consist of almost 66 million tokens written from 1989–

[2] Ološtiak and Ološtiaková [6, p. 230] mention as many as 38 suffixes, though 25 sufixes in their sample – derived from the *Slovník koreňových morfém slovenčiny* comprising 66,500 analysed lexical units – represent only 1% of adjectives.

2014. It comprises more than 80% of thematic journals and newspapers. Similarly, specialised corpus ecn-2.0-public [12], devoted to the field of economics, includes as much as 96.24% of specialised texts published in thematic journals and newspapers. Texts of this corpus come from 1992–2014 and comprise almost 165 million tokens.

3 THEORETICAL BACKGROUND AND METHODOLOGY

Morphological productivity represents one of the most contentious linguistic issues and is the focus of extensive research and discussions.[3] One complex theory of word-formation productivity was presented by a Czech linguist, Miloš Dokulil, in his work *Teorie tvoření slov* in 1962 [13], which also included the differentiation of the *systemic* and the *empirical productivity* (also termed 'parole' or 'real productivity' [14]), the latter determined by extra-linguistic factors. With the availability of extensive corpora, it is possible to measure, identify, and analyse the concept of Dokulil's empirical productivity of a word-formation type or element in a language at a given time. Dokulil believed that even "approximative data concerning the quantitative use of a given word-formation process or element are of paramount importance in the overall picture of a given language in general and for the characteristics of its lexicon in particular" [Ibid., p. 77]. Moreover, Dokulil's theory is also inspirational in the differentiation of the so-called *absolute frequency* of word-formation processes, types, and elements and the *relative frequency*, which is register- or domain-dependent.

Echoing Dokulil's theory of productivity, some contemporary Czech linguists elaborate on and verify his assumptions on corpus data (see, e.g., [14], [15], [16], [17] or [18]). Štícha advocates the need to analyse suffixes one by one, to identify the frequency and ratio of derived words, coined with these suffixes, in a given corpus [14, p. 100]. Štícha proposes the analysis of empirical/parole productivity not only on big corpora, but also on a series of corpora of different size and composition [Ibid., p. 104]. Corpus findings and statistical data indicate that empirical/parole productivity could be differentiated further into general and specialised categories which will be of special interest in the context of this study [Ibid.].

The main "trend" concerning statistical-based research of productivity, introduced especially by the studies of Harald Baayen and his colleagues in the 1990s ([19], [20], [21], [22], [23]), claims an importance for *hapax legomena* in a given corpus in determining the degree of productivity of a word-formation type or element. The rationale behind this method is that lemmas which occur only once in

[3] More information can be found, e.g., in Hulse, V.: Productivity in morphological negation: a corpus based approach. The University of Manchester (2011). Available at: https://www.research.manchester.ac.uk/portal/en/theses/productivity-in-morphological-negation-a-corpusbased-approach(266d2241-a266-4b99-8fab-e19571381d8f).html.

a corpus are indicative of the creation of new words. However, many researchers criticize the significance attributed to hapaxes in measuring morphological productivity and emphasize the fact that not all hapaxes represent new coinages, on the contrary, this group often consists of peripheral lexical units including archaisms or words that simply happen to occur only once in a given corpus. To do Baayen justice, he explicitly states that hapaxes only correlate to the number of neologisms and that "they only function as a tool for a statistical *estimation method* aimed at gauging the rate of expansion of morphological categories" [24, p. 906].

In his 2009 paper [24], Baayen proposed a more elaborate theory of (morphological) productivity distinguishing three levels:

1) realized productivity, which reflects the productivity of a word-formation type/process in the past and which can be "estimated by the type count", i.e., the number of lemmas in a corpus;

2) expanding productivity, which is an estimate of the contribution of a morphological category to the growth rate of the total vocabulary. Baayen suggests it be calculated as the ratio of hapaxes with affix X/all hapaxes in a corpus;

3) potential productivity, which enables one to "estimate the growth rate of the vocabulary of the morphological category itself" and can be calculated as the number of hapaxes with affix X/tokens with affix X. Baayen used this method already in [19], stressing that it is to show only the statistical probability ratio of future coinages.

Due to legitimate criticism of hapax significance mentioned earlier, this analysis is based on manually *cleaned-up* data. Lists of hapaxes extracted from every corpus and subcorpus were checked and several groups of lemmas excluded:

a) lemmas of Slovak origin including incidentally the same sequence of characters as the analysed suffixes (exclusion not only from the list of hapaxes, but also from the list of lemmas with the suffix X);

b) lemmas with typos or orthographical mistakes;

c) lemmas found in general Slovak dictionaries and the Dictionary of Foreign Words (most of them representing terms of specialised domains);

d) lemmas found in two most extensive SNC corpora: exclusion of lemmas with 3 and more occurrences in general corpus prim-9.0-juls-all [9] and in legal corpus legal-1.1 [10], provided that those occurrences come from 3 different sources and 3 different years.

Overall, the exclusion ranged from 26% up to 100% of hapaxes in individual (sub)corpora. However, the cleaned-up lists of hapaxes may still comprise lemmas that are not neologisms, due to the lack of up-to-date specialized dictionaries and the extent and content of the corpora used.

Moreover, if a list of lemmas with suffix X comprised two lemmas differing only in the usage or non-usage of a hyphen, these were merged, as well as lemmas (not being proper names) with a capitalised and non-capitalised first letter.

For practical reasons, a thorough manual check and clean-up of lists of all hapaxes from big corpora is more than time-consuming and unfeasible. Therefore, further analyses will focus only on Baayen's realized and potential productivity:

1. **realized productivity** in order to determine the productivity of suffixes with respect to past and present linguistic situations, completed with the corpus statistics reflecting their usage;

2. **potential productivity** in order to estimate the rate at which new types are to be expected to appear. However, I decided not to use the hapax/token method, but the hapax/type method in line with Van Marle's reasoning [25] that token frequency is not as relevant a variable in the measure of productivity as the number of lemmas. Moreover, I assume that an estimation of future productivity should be related to contemporary and past productivity.

As far as the corpus search is concerned, I did not base the queries on lemmatization and morphological tagging of the (sub)corpora because with Latinate words the lemmatization and tagging proved to be inadequate and erroneous. Therefore, I opted for simple search of specific ending of a token, e.g. [lemma=".*álny"], combined with automatic filtering of words with incidentally the same string of characters.

4 RESEARCH RESULTS

4.1 Usage of suffixes and realized productivity

Table 2 shows the raw frequency of tokens with analysed suffixes occurring in 6 selected (sub)corpora. For the sake of comparison, the second column features the normalised frequency (ipm) of these tokens and, thus, enables an inference of their usage in general and specialised domains. The suffixes have been listed in order of decreasing number of tokens, which is mirrored by the decreasing normalised frequency, except for the order of the pair *-ózny* and *-órny* in legal corpus and *-árny* and *-itný* in the economic corpus. Suffix *-álny* is clearly the most widely used in all (sub)corpora, while *-órny* is at the opposite end of the frequency axis in 5 (sub)corpora. Four out of five suffixes clearly reach higher normalised frequencies in the general corpus of specialised and academic texts (prim-9.0-public-prf) compared to reference corpus, as expected. Frequency differences between reference corpus and prim-9.0-public-prf, as well as those between prim-9.0-public-prf and each specialised (sub)corpus, were subjected to a test of significance test (log likelihood test), which confirmed the significance of observed statistical data with the exception of the suffix *-árny* in medical texts and the suffix *-órny* in legal texts. It is also noteworthy that normalised frequencies of suffixes in medical subcorpus equal or considerably exceed the ipm in prim-9.0-public-prf while the ipm of suffixes in religious corpus is manifestly lower than in the reference corpus.

suffix	prim-7.0-frk	prim-9.0-public-prf	prim-9.0-juls-all (MED)	legal-1.1.ver. bz 1991_2011	blf-2.0	ecn-2.0-public
	Tokens	Tokens	Tokens	Tokens	Tokens	Tokens
	IPM	IPM	IPM	IPM	IPM	IPM
-álny	547,742	537,682	31,408	102,863	128,472	597 005
	2163.81	3594.57	4423.94	3061.38	1948.9	3618.5
-árny	84,780	94,586	4,529	25,980	14,253	38,178
	334.92	634.14	637.93	773.21	216.22	231.40
-itný	44,655	30,402	2,683	2,256	9,176	52,574
	176.41	203.25	377.91	67.14	139.20	318.66
-ózny	16,531	8,863	1,416	482	2,321	9,440
	65.30	59.25	199.45	14.35	35.21	57.22
-órny	4,672	4,537	884	1,959	620	2,039
	18.46	30.33	124.51	58.30	9.41	12.36

Tab. 2. Frequency and normalised frequency of analysed suffixes in the respective (sub)corpora

The estimate of realized productivity, or the insight into the extent of past new coinages by means of analysed suffixes, can be seen in table 3, which shows not only the number of lemmas in the selected (sub)corpora, but also normalised counts of lemmas per million. Again, the first position in the table is taken by the suffix *-álny* and the last place by the suffix *-órny*. Both absolute and normalised counts of lemmas are distinctly higher in prim-9.0-public-prf compared to the reference corpus, and from among specialised (sub)corpora, it is the medical field in which all five suffixes appeared most frequently. Only the number of lemmas in economic corpus, with the exception of lemmas with *-itný*, is closer to the reference corpus than to the corpus of specialised texts. Suffixes *-itný* and *-ózny* feature very similar statistics with the exception of medical texts, in which lemmas with *-ózny* are twice as numerous per million than those with *-itný*. The same reversed order of these two suffixes, compared to other corpora, is in the reference corpus.

suffix	prim-7.0-frk		prim-9.0--public-prf		prim-9.0-juls-all (MED)		legal-1.1.ver. bz 1991_2011		blf-2.0		ecn-2.0-public	
-álny	1,006	3.97	1,179	7.88	519	73.10	322	9.58	643	9.75	753	4.56
-árny	298	1.18	366	2.45	180	25.35	100	2.98	144	2.18	204	1.24
-itný	90	0.36	97	0.65	36	5.07	41	1.22	55	0.83	118	0.72
-ózny	115	0.45	119	0.80	72	10.14	40	1.19	52	0.79	83	0.50
-órny	25	0.09	35	0.23	13	1.83	10	0.30	20	0.30	27	0.16

Tab. 3. Number of lemmas in a given (sub)corpus followed by the same count normalised per million tokens

To complete the picture, the last table in this part presents a statistical method frequently used in analyses of (morphological) productivity – type/token ratio of analysed suffixes. These data in table 4 reflect both the existence of types with one of the suffixes, as well as the extent to which these types are used. In this two-dimensional perspective, the ranking of analysed suffixes is reversed compared to the previous table – the first place is occupied either by *-ózny* or *-órny,* while *-álny* can be found at the bottom of table in 5 (sub)corpora. However, it is not possible to compare these ratios across the corpora as they were calculated from raw token frequencies and counting of lemmas.

Type/token ratio											
prim-7.0-frk		prim-9.0-public-prf		prim-9.0-juls-all (MED)		legal-1.1.ver.bz 1991_2011		blf-2.0		ecn-2.0-public	
-ózny	0.006956627	-ózny	0.013426605	-ózny	0.050847458	-ózny	0.082987552	-órny	0.032258065	-órny	0.013241785
-órny	0.005351027	-órny	0.007714349	-árny	0.039743873	-itný	0.018173759	-ózny	0.022404136	-ózny	0.008792373
-árny	0.003514980	-árny	0.003869494	-álny	0.016524452	-órny	0.005104645	-árny	0.010103136	-árny	0.005343391
-itný	0.002015452	-itný	0.003190580	-órny	0.014705882	-árny	0.003849115	-itný	0.005993897	-itný	0.002244455
-álny	0.001836631	-álny	0.002192746	-itný	0.013417816	-álny	0.003130377	-álny	0.005004982	-álny	0.001261296

Tab. 4. Type/token ratios in a given (sub)corpus

4.2 Potential productivity

As I indicated in part 3, for calculating the potential productivity of analysed suffixes, I decided to use not the hapax/token method, but the hapax/type method, in order to emphasize the relation of hapaxes to types rather than tokens. Table 5 introduces the potential productivity ratios in decreasing order. Note first that the order of suffixes in the reference corpus and prim-9.0-public-prf overlaps only partially: while *-árny* and *-álny* keep the same 3rd and 4th place, respectively, *-órny* occupies the last place in the reference corpus, but tops prim-9.0-public-prf. Similarly, suffix *-ózny* takes the first place in the reference corpus, but the 4th place in prim-9.0-public-prf. If we compare the situation in prim-9.0-public-prf and specialised (sub)corpora, the most productive suffix seems to be *-órny,* though in medical texts, it occupies the last place of the ranking and the 3rd place in economic texts. Very different ranking can be observed for the suffix *-ózny*: 4th place in prim-9.0-public-prf and the religious corpus, 2nd place in the legal and economic corpora and 1st place in medical texts, as expected. Suffix *-itný* is either least productive (in prim-9.0-public-prf, legal and religious corpora) or the most productive (1st place in economic and 2nd place in medical texts). Relatively stable potential productivity is manifested by the suffix *-álny*: 3rd place in the ranking of three (sub)corpora, 2nd place in religious and 4th place in economic texts. The last but not least, suffix *-árny* has the 2nd highest productivity in prim-9.0-public-prf, 4th in medical and legal texts, 3rd in religious texts, but lowest in economic texts.

Hapax/type ratio												
prim-7.0-frk		prim-9.0-public-prf		prim-9.0-juls-all (MED)		legal-1.1.ver.bz 1991_2011		blf-2.0		ecn-2.0-public	
-ózny	0.252173913	-órny	0.228571429	-ózny	0.208333333	-órny	0.1	-órny	0.2	-itný	0.237288136
-árny	0.201342282	-árny	0.210382514	-itný	0.138888889	-ózny	0.075	-álny	0.188180404	-ózny	0.228915663
-álny	0.173956262	-álny	0.209499576	-álny	0.129094412	-álny	0.052795031	-árny	0.145833333	-órny	0.185185185
-itný	0.122222222	-ózny	0.201680672	-árny	0.1	-árny	0.05	-ózny	0.134615385	-álny	0.177954847
-órny	0.12	-itný	0.195876289	-órny	0.076923077	-itný	0	-itný	0.090909091	-árny	0.142156863

Tab. 5. Hapax/type ratios in a given (sub)corpus. The suffixes in each subtable are listed in order of their decreasing ratio

In order to put the data in one more perspective, let us regroup the (sub)corpora in the order of their increasing corpus size and show the ratio of hapaxes as a percentage. The aim of the reordering is to answer the question of František Štícha [14, p. 255] regarding what the ratio of hapaxes will be with the increasing size of corpora. Štícha hypothesized that a significant increase of lemmas with a specific suffix correlated with an increase of low-frequency lemmas with the same affix would testify to/indicate a high real productivity of this type in a given time. However, our (sub)corpora differ significantly not only in terms of size but also in terms of types of texts and their proportion, therefore, it is not possible to test this hypothesis fully, just to indicate discernible trends. Table 6 shows that apart from the reference corpus data, two corpora – legal and economic types – are incoherent with the increasing number of either lemmas or hapaxes, or both. We can hypothesize that the reason lies in their composition and, possibly, in the character of the discipline, e.g., legal domain is rather hesitant towards linguistic innovations. Only in the case of two suffixes, *-áln*y and *-órny,* it is possible to observe both the increase of lemmas and ratio of hapaxes in at least three (sub)corpora – medical subcorpus, religious corpus and prim-9.0-public-prf. An interesting drop in data can be seen between the medical subcorpus and the legal corpus, which is almost ten times larger) – except for the number of *-itný* lemmas and the ratio of *-órny* hapaxes. Similar drop, including the exception of *-itný* lemmas and hapaxes, is between prim-9.0-public-prf data and the data from economic texts. If we narrow our focus to the difference between general and specialised texts, we can observe that the number of lemmas, as well as the hapax ratio in prim-9.0-public-prf, is higher compared to the reference corpus, except for the ratio of *-ózny* hapaxes.

suffix	prim-9.0-juls-all (MED)		legal-1.1.ver. bz 1991_2011		blf-2.0		prim-9.0--public-prf		ecn-2.0-public		prim-7.0-frk	
-álny	519	12.91%	322	5.28%	643	18.82%	1179	20.95%	753	17.8%	1006	17.4%
-árny	180	10%	100	5%	144	14.58%	366	21.04%	204	14.22%	298	20.13%
-itný	36	13.89%	41	0%	55	9.09%	97	19.59%	118	24.14%	90	12.22%
-ózny	72	20.83%	40	7.5%	52	13.46%	119	20.17%	83	22.89%	115	25.22%
-órny	13	7.69%	10	10%	20	20%	35	22.86%	27	18.52%	25	12%

Tab. 6. Ratio of hapaxes in % in (sub)corpora ordered from the smallest (medical subcorpus) to the biggest corpus (prim-7.0-frk)

To sum it up, the most productive suffix in terms of past coinages is *-álny*, especially in medical, legal and religious texts, the least productive, in this perspective, is *-órny,* though the number of different lemmas with this suffix is significantly higher in prim-9.0-public-prf and specialised (sub)corpora compared to the reference corpus.

From a future productivity point of view, the picture is less clear-cut: the most productive suffix is *-órny* but only in three corpora. In the remaining three (sub) corpora, the ranking is topped in two instances by *-ózny* and once by *-itný*. Moreover, the same suffix *-órny* seems to be least productive in future regarding general and medical texts. In as many as three (sub)corpora, it is the suffix *-itný* that has taken the final place in the productivity ranking. Similarly, the last place in economic texts productivity ranking is occupied by the suffix *-árny*.

5 CONCLUSION

A proposed analysis of word-formation productivity of selected suffixes in Slovak indicates noteworthy differences depending on domains and registers. However, these differences need to be verified in specialised corpora, balanced in terms of genres and types. An open question remains as to whether the analysis of the share of neologisms in low-frequency lemmas would not change the overall picture, as several researchers note that word-frequency distribution of productive affixes is supposed to be distinctly shifted towards low-frequency lemmas comprising new coinages.

ACKNOWLEDGEMENTS

The paper has been written within the Slovak National Corpus project supported by the Slovak Academy of Sciences, Ministry of Education, Science, Research and Sport of the Slovak Republic, Ministry of Culture of the Slovak Republic and the Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences.

References

[1] Horecký, J. (1999). Internacionalizácia a europeizácia slovenčiny. In Internacionalizácia v súčasných slovanských jazykoch. (Ed.) J. Bosák, pages 80–82, Bratislava, Veda.

[1] Buzássyová, K., Horecký, J., Bosák, J. et al. (1989). Dynamika slovnej zásoby súčasnej slovenčiny. Bratislava, Vydavateľstvo Slovenskej akadémie vied.

[2] Jedlička, A. (1974). Spisovný jazyk v současné komunikaci. Praha, Univerzita Karlova.

[3] Buzássyová, K. (1991). Opakovaná internacionalizácia a problém identifikácie morfologických a lexikálnych jednotiek. Jazykovedný časopis, 42(2), pages 89–104.

[4] Bozděchová, I. (2009). Současná terminologie (se zaměřením na kolokační termíny z lékařství). Praha, Karolinum.

[5] Ološtiak, M., and Ološtiaková, L. (2015). Formálno-procesuálne aspekty slovotvornej motivácie. In Kvalitatívne a kvantitatívne aspekty tvorenia slov v slovenčine, pages 207–308, Prešov, Filozofická fakulta Prešovskej univerzity v Prešove.
[6] Slovenský národný korpus. Korpus prim-7.0-frk. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2018. Accessible at: https://korpus.sk.
[7] Slovenský národný korpus. Korpus prim-9.0-public-prf. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2020. Accessible at: https://korpus.sk.
[8] Slovenský národný korpus. Korpus prim-9.0-juls-all. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2020. Accessible at: https://korpus.sk.
[9] Slovenský národný korpus. Korpus legal-1.1. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2013. Accessible at: https://korpus.sk.
[10] Slovenský národný korpus. Korpus blf-2.0. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2014. Accessible at: https://korpus.sk.
[11] Slovenský národný korpus. Korpus ecn-2.0-public. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2016. Accessible at: https://korpus.sk.
[12] Dokulil, M. (1962). Teorie tvoření slov. Praha, Nakladatelství ČSAV.
[13] Štícha, F. (2012). Jak v epoše elektronických korpusů následovat Miloše Dokulila (Miloši Dokulilovi ke stému výročí narození). In Jazykovědné aktuality, 49, pages 95–107.
[14] Štícha, F. (2002). K Dokulilovu pojmu slovotvorné produktivity (z hlediska korpusové analýzy). In Čeština doma a ve světě, 4, pages 302–310.
[15] Štícha, F. (2007). Korpusové statistiky a slovotvorná produktivita. In Grammar & Corpora/ Gramatika a korpus 2005. F. Štícha and J. Šimandl (eds.). Praha, Academia, pages 250–257.
[16] Štícha, F. (2009). Slovotvorná produktivita a gramatičnost: gradační expresivní adjektiva s prefixy pra-, pře- a vele- v současné psané češtině. In Eslavística Complutense, 9, pages 145–170.
[17] Ševčíková, M. (2014). Zjišťování slovotvorné produktivity z korpusových dat: přípony odvozující názvy vlastností. In Naše reč, 97, pages 228–240.
[18] Baayen, H. R., and Lieber, R. (1991). Productivity and English derivation: a corpus-based study. In Linguistics, 29, pages 801–843.
[19] Baayen, H. R. (1992). Quantitative aspects of morphological productivity. In Yearbook of Morphology 1991. G. E. Booij and J. van Marle (eds.), pages 109–149, Dordrecht, Kluwer Academic Publishers.
[20] Baayen, H. R. (1993). On frequency, transparency, and productivity. In Yearbook of Morphology 1992. G. E. Booij and J. van Marle (eds.), pages 181–208, Dordrecht, Kluwer Academic Publishers.
[21] Baayen, H. R. (1994). Productivity in production. In Language and Cognitive Processes, 9, pages 447–469.
[22] Baayen, H. R., and Renouf, A. (1996). Chronicling The Times: Productive Lexical Innovations in an English Newspaper. In Language, 72, pages 69–96.
[23] Baayen, R. H. (2009). Corpus linguistics in morphology: morphological productivity. In Corpus Linguistics. An international handbook. A. Lüdeling and M. Kyto (eds.), pages 900–919, Berlin, Mouton De Gruyter.
[24] van Marle, J. (1992). The relationship between Morphological Productivity and Frequency: A Comment on Baayen's Performance-Orientated Conception of Morphological Productivity. In Yearbook of Morphology 1992. G. E. Booij and J. van Marle (eds.), pages 151–163.

sciendo
DOI 10.2478/jazcas-2021-0037

THE MENZERATH-ALTMANN LAW AS THE RELATION BETWEEN LENGTHS OF WORDS AND MORPHEMES IN CZECH

KATEŘINA PELEGRINOVÁ[1] – JÁN MAČUTEK[2,3] – RADEK ČECH[1]

[1] Department of Czech Language, University of Ostrava, Ostrava, Czech Republic
[2] Mathematical Institute, Slovak Academy of Sciences, Bratislava, Slovakia
[3] Department of Mathematics, Constantine the Philosopher University, Nitra, Slovakia

PELEGRINOVÁ, Kateřina – MAČUTEK, Ján – ČECH, Radek: The Menzerath-Altmann law as the relation between lengths of words and morphemes in Czech. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 405 – 414.

Abstract: It is shown that the mean morpheme length (measured in phonemes) decreases with the increasing length of word types (in morphemes) in Czech texts, i.e., these language units behave according to the Menzerath-Altmann law. The law is not valid in general for word tokens. Some hints towards an interpretation of parameters are presented.
Keywords: Menzerath-Altmann law, word, morpheme, phoneme, Czech

1 INTRODUCTION

The Menzerath-Altmann law [1] (henceforward MAL) is, together with the Zipf law [2] and the law of brevity [3] (which was, in fact, formulated also by Zipf), one of the best known laws in quantitative linguistics. Its special case was first articulated by Paul Menzerath [4] who studied the German vocabulary and observed that longer words consist, on average, of shorter syllables. The law was later substantially generalized by Gabriel Altmann [5]. The current version of the MAL claims that greater constructs consist on average of smaller constituents, with constructs and constituents being neighbours in the language unit hierarchy (such as, e.g., words and syllables, sentences and clauses, etc.). Sometimes even a more general form is used, namely, the size of the construct is a function of the mean size of its constituents. The function does not necessarily have to be strictly decreasing. It can increase to its peak (achieved usually for constructs of size two) and decrease from the peak to the right, see theoretical considerations in [6, p. 7] and examples in [5, p. 8; Table 4] or [6, p. 54; Table 5.7]. The general mathematical expression for the MAL is

$$y(x)=a\,x^{b}\,\mathrm{e}^{-cx}, \tag{1}$$

with $y(x)$ denoting the mean size of constituents in constructs of size x; a, b, and c are parameters of the model. A special case of this general formula, namely

(2) $y(x)=a\,x^{b}$,

is very often sufficient to fit data. This simpler version is appropriate only if the mean constituent size achieves its maximum for constructs of length 1, and the mean constituent size then decreases with the increasing construct size.

The validity of the MAL was corroborated for several language levels and in many languages. As examples, we mention the MAL as the model for the relation between lengths of words (in syllables) and syllables (in graphemes) [7], canonical word forms (in syllables) and syllables (in phonemes) [8], word length motifs (in words) and words (in syllables) [9], sentences (in clauses) and clauses (in words) [10]. We note that while there seems to be a consensus on the hierarchy of low-level language units (speech sound/phoneme – syllable/morpheme – word)[1], the situation from word higher on is not so clear. Until relatively recently, clause was considered the "upper neighbour" of word (see e.g. [10]), but a rapid development of computational methods in syntax made it possible to take into account also other, intermediate level units (e.g., the MAL was shown to be valid for the relation between the lengths of clauses and syntactic phrases which are directly dependent on the clause predicate, see [11]). In addition, new, "non-traditional" units (different kinds of motifs, see an overview in [12]) were defined, which follow the MAL as well.

The first attempt to study the MAL specifically as a model for the relation between lengths of words (in morphemes) and morphemes (in phonemes) can be found in [13], where 15,011 German words' lemmas from a dictionary are analysed. A similar approach was chosen in [14] – the author uses databases of morphologically segmented word lemmas from Dutch, English, and German, which consist of 124,136, 52,447, and 50,708 word lemmas, respectively. Here, morpheme length is measured both in phonemes and graphemes. Both papers (and both choices of units in which morpheme length was measured in the latter one) result in data which can be modelled by the MAL in form (2), i.e., it holds that the longer the word, the shorter the mean length of its morphemes. The same is true for a short Turkish text [15, p.20], with morpheme length measured in the number of phonemes.[2] Both word types and tokens from a Czech novella were taken into account in [17]. A decreasing trend of morpheme lengths can be observed, but the two curves differ (the one for types is steeper).

[1] Needless to say, there are some (mainly methodological) problems related also to the analysis of these low-level units, such as e.g., their different definitions, the status of zero-syllable prepositions, etc.

[2] In [15], it is not specified whether word types or tokens are analysed. In addition, word length in syllables is studied in the same text in [16] but, curiously enough, the total numbers of words differ in the two works (559 words in Text 7 from [15], and 587 words in Text 2 from [16]). The difference between these two numbers is too small to be explained by different approaches, i.e., as the number of types in [15] and the number of tokens in [16]. In such case, the type-token ratio would achieve an extremely implausible high value of 0.95.

This paper presents the first systematic study of the relation between word length and morpheme length, performed on the sample of 15 Czech texts which are morphologically segmented by the same method. Function (1) is a good model for 14 of the texts, the only one for which the fit of the model is not sufficiently good is the shortest text in the sample. Some hints towards an interpretation of the parameters can be found in the conclusion.

2 METHODOLOGY AND DATA

The morphological segmentation applied in this paper is based on the retrograde morphemic dictionary of Czech [18]. However, [18] contains only dictionary forms of words (i.e. lemmas), and since Czech is an inflectional language, also rules for inflected word forms are needed. Moreover, there are several groups of words (proper names, pronouns, particles, interjections) which require a special approach.

The segmentation of inflected word forms of nouns, adjectives, numerals, and verbs mostly follows the Czech grammar [19]. The segmentation rules from [19] were modified or specified more in detail as follows.

1. For nouns and adjectives, markers of grammatical cases were reconsidered in order to avoid a high degree of allomorphy. For example, the morpheme *ch* serves as the marker of the locative plural in this paper, while according to [18], the locative plural is marked by six allomorphs: *-ech*, *ích*, *-ách*, *-ich*, *ěch*, *-ch*.[3] The vowel which precedes the morpheme *-ch* in the locative plural (if there is any) is considered a separate morpheme (in [20], it is called a connecteme). Other grammatical cases are treated analogously.
2. Czech pronouns can be inflected and, at the same time, they constitute a closed class containing a limited number of words. Therefore, there are less options for inter-paradigmatic comparisons (within this class as well as with other nominal parts of speech, i.e., nouns and adjectives). Consequently, the segmentation rules for pronouns found in grammars are often more ambiguous than for other parts of speech. Led again by the motivation to reduce allomorphy, we decided to apply a deeper (i.e., more detailed) morphological segmentation. We demonstrate our approach on the example of the instrumental plural form *našimi* of the first person plural possessive pronoun *náš* 'our'. First, the inflectional morpheme is separated: *naš-imi*. Then, applying the above mentioned rules, *-mi* is considered the marker of the instrumental plural, and *-i-* which precedes it a connecteme: *naš-i-mi.* But in Czech there is also the instrumental plural affix *-ma*[4], contrasting

[3] These allomorphs occur in nouns, adjectives, and some pronouns.

[4] The affix *-ma* originally marked the dual number. In contemporary Czech it is used with nouns referring to paired body parts (and in forms agreeing with such nouns).

with *-mi* (*našim-a* vs. *našim-i*). This contrast leads to the segmentation into *našim-i.* Next, the inter-paradigmatic comparison with the second person plural possessive pronoun *váš* 'your' results in the segmentation *n-aš-i-m-i*, as these two pronouns contrast only in their consonant roots *n-* vs. *v-*. The final step is carried out on the basis of inter- and intra-paradigmatic comparisons of inflected forms of the possessive pronouns *náš* 'our', *váš* 'your' with the corresponding (non-possessive) personal pronouns *my* 'we', *vy* 'you'. Pronouns *my* and *vy* have stems *ná-* and *vá-*, respectively, in all cases with the exception of the nominative form. Therefore the final segmentation is *na-š-i-m-i.*

3. In proper names, only inflectional affixes and productive derivational suffixes are segmented. The remaining stems are not analysed morphologically here (e.g., nominative singular *Prah-a* 'Prague', genitive singular *Prah-y*; *Pelegrin-ov-á* – the female version of a surname with the suffix *-ov-*; *Fin-sk-o* 'Finland' etc).
4. Particles and interjections in Czech are uninflected parts of speech with an ambiguous delimitation and no consensus concerning their morphological segmentation. We only segment those particles and interjections which are morphologically transparent.

The texts under analysis were processed semiautomatically. Word forms were morphologically segmented manually at their first occurrence. Thus, a dictionary of morphologically segmented word forms was created. Then, a computer script[5] was written, which found the segmentation in the dictionary at further occurrences of the word forms.[6]

The texts[7] which serve as our language material were taken from a corpus of works by Czech writer Karel Čapek (the corpus described in [25]). In particular, we took five short stories (denoted as S1 – S5 below), five personal letters (L1 – L5), and five studies on philosophy (P1 – P5).[8] The texts were transcribed in such a way that the number of letters is equal to the number of phonemes the word consists of (e.g. *hoch* 'boy' is transcribed as *hox*).

[5] The script is written in Python. It is available upon request.

[6] The dictionary was enlarged every time the program encountered a hitherto unsegmented word form.

[7] Within our research framework, we prefer to work with texts rather than with corpora. At the same time, we are convinced that both of these approaches are reasonable, and both have their own advantages as well as limitations. See e.g. [21], [22], [23], and [24] for text vs. corpus discussions and related topics.

[8] The following texts were chosen: the first five short stories from the collection *Povídky z druhé kapsy*, personal letters with numbers 749, 753, 755, 756, and 761 (as they are numbered in https://search.mlp.cz/cz/titul/korespondence/43837/#book-content), and chapters one, two, three, four, and nine from the study *Pragmatismus*.

3 RESULTS

The mean morpheme lengths (expressed in the number of phonemes) in word types of particular lengths (counted in morphemes) which occur in the texts under analysis can be found in Tables 1, 2, and 3. As the mean is strongly affected by extreme values if the sample size is small, we decided to pool the data so that the minimum frequency in each category is ten.[9] In the pooled categories, word length is represented by the weighted arithmetic mean of the pooled words, with length frequencies serving as the weights.[10] In the tables, *WL* denotes word length, *MML* the mean morpheme length in words consisting of a particular number of morphemes, and *fr* the frequency with which particular word lengths occur in the texts.

The data were fitted to the MAL in form of (2), see Section 1. The fitness of the model was evaluated in terms of the determination coefficient R^2. The fit is usually considered satisfactory if $R^2>0.9$, see [27]. As in (2) it holds $y(1)=a$, the value of the parameter *a* was set as the mean length (in phonemes) of monomorphemic words in particular texts. We thus follow the approach applied to the relation between word length and syllable length from [7]. The value of the parameter *b* maximizes the determination coefficient.[11] The values of the two parameters and R^2 are also presented in the tables.

	S1		S2		S3		S4		S5	
WL	*MML*	*fr*	*MML*	*fr*	*MML*	*fr*	*MML*	*fr*	*MML*	*fr*
1	4.30	145	3.82	117	3.61	115	3.76	159	3.65	127
2	2.39	261	2.41	213	2.36	200	2.46	342	2.33	260
3	1.92	259	1.89	210	1.92	244	1.94	400	1.88	270
4	1.80	199	1.76	155	1.76	131	1.75	265	1.78	171
5	1.73	66	1.71	57	1.74	47	1.64	98	1.70	56
6					1.53	13				
6.19									1.50	16
6.21							1.63	28		
6.35			1.60	20						
6.42	1.61	26								
a	4.30		3.82		3.61		3.76		3.65	
b	-0.64		-0.55		-0.51		-0.54		-0.53	
R^2	0.941		0.959		0.973		0.973		0.968	

Tab. 1. Word length (in morphemes) and morpheme length (in phonemes): short stories, word types

[9] The minimum frequency of ten is only a rule o thumb, see e.g., [11] and [26]. Another possibility is to neglect the lengths with too low frequencies, see, e.g., [8].

[10] We demonstrate the pooling on the example of the first short story, see Table 1. There are 19 words of length six, three words of length seven, and four words of length eight. Therefore, these word lengths are pooled into one category, with the weighted mean being $(19\times6+3\times7+4\times8)/(19+3+4)=6.42$. The mean morpheme length is evaluated as the mean length of morphemes in all words from this category.

[11] The values of the parameter *b* were determined using NLREG software (www.nlreg.com).

	L1		L2		L3		L4		L5	
WL	*MML*	*fr*	*MML*	*fr*	*MML*	*fr*	*MML*	*fr*	*MML*	*fr*
1	3.05	81	2.98	49	3.17	63	2.64	47	3.12	42
2	2.18	132	2.32	79	2.07	107	2.13	82	2.03	76
3	1.85	158	1.83	65	1.81	101	1.85	82	1.76	69
4	1.77	108	1.86	45	1.82	70	1.74	37	1.62	36
5	1.62	36							1.73	25
5.18							1.58	17		
5.23					1.74	40				
5.24			1.61	25						
6.17	1.73	12								
6.20									1.76	10
a	3.05		2.98		3.17		2.64		3.12	
b	-0.39		-0.38		-0.44		-0.31		-0.42	
R^2	0.934		0.976		0.899		0.998		0.833	

Tab. 2. Word length (in morphemes) and morpheme length (in phonemes): letters, word, types

	P1		P2		P3		P4		P5	
WL	*MML*	*fr*	*MML*	*fr*	*MML*	*fr*	*MML*	*fr*	*MML*	*fr*
1	3.25	60	3.07	46	4.17	143	3.68	90	3.00	46
2	2.47	111	2.25	62	2.70	184	235	205	2.28	66
3	2.07	121	2.03	62	2.18	165	1.98	232	2.00	82
4	1.93	111	1.94	51	1.99	143	1.89	220	1.86	49
5	1.84	79	1.79	41	1.85	90	1.71	142	1.71	23
6	1.85	21					1.73	61		
6.36					1.85	39				
6.52			1.75	27						
7									1.53	11
7.14	1.74	14					1.57	28		
a	3.25		3.07		4.17		3.68		3.00	
b	-0.35		-0.34		-0.52		-0.48		-0.35	
R^2	0.967		0.955		0.965		0.947		0.995	

Tab. 3. Word length (in morphemes) and morpheme length (in phonemes): studies on philosophy, word types

The MAL expressed by formula (2) fits the relation between word length in morphemes and the mean morpheme length in phonemes very well for 13 out of 15 texts under analysis. The fit for text L3 is practically on the threshold of 0.9. The only exception with a low value of R^2 is text L5. However, this text contains only 258 word types (it is the shortest in our sample), and, moreover, there are exactly ten

words (i.e., the minimum which is accepted) in the pooled category with the longest words. If this criterion is made stricter and words with length five or more are pooled into one category, the fit becomes acceptable, with $R^2=0.91$. The relation between word length and morpheme length is demonstrated also in Figure 1, where data for text P1 (see Table 3) are used.

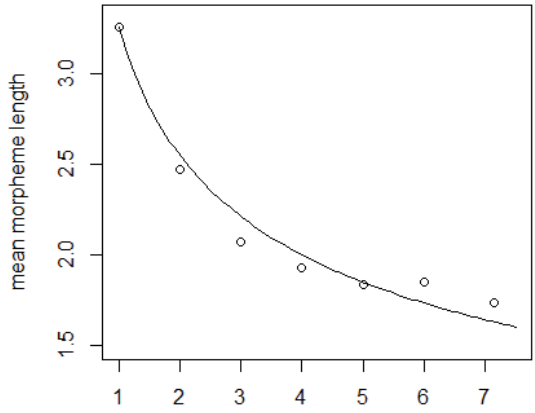


Fig. 1. The MAL in text P1

One can see a different pattern of behaviour if word tokens (as opposed to word types) are analysed (see data for short stories in Table 4). In some texts, the relation between the mean morpheme length fluctuates, and one cannot speak about a decreasing trend in general. In general, the MAL in the form given by (2) fails to achieve an acceptable fit.[12]

	S1		S2		S3		S4		S5	
WL	*MML*	*fr*	*MML*	*fr*	*MML*	*fr*	*MML*	*fr*	*MML*	*fr*
1	2.44	312	2.33	109	2.74	235	2.27	122	2.31	118
2	1.90	273	2.11	118	1.83	184	1.94	143	1.81	132
3	1.72	229	1.78	80	1.77	138	1.77	105	1.62	108
4	1.74	121	1.84	50	1.80	74	1.74	37	1.59	42
5	1.57	40								
5.20					1.72	45				
5.22							1.61	18		
5.23			1.61	26						
6.15	1.72	13								
6.17									1.80	12

Tab. 4. Word length (in morphemes) and morpheme length (in phonemes): short stories, word tokens

[12] The MAL in the form given by (1) fits the data for word tokens very well. However, using this formula would mean fitting roughly five or six data points with a function with three uninterpreted parameters.

This behaviour can be explained – admittedly, only speculatively for the time being – as a display of a competition between two "language forces" represented by the MAL on the one hand, and by the Zipf law of brevity on the other. According to the latter, shorter units are preferred. If the law of brevity is valid also within words of particular lengths (e.g., if words consisting of three shorter morphemes occur more often than words with three longer morphemes), the MAL may (see, e.g., [17]) or may not (see, e.g., text S1 in Table 4) hold for word tokens, depending on how strongly the law of brevity prefers shorter morphemes.

4 CONCLUSION

The presented results corroborate the validity of the MAL on new language material. In addition, we are able to provide some hints towards an interpretation of the parameters of the model.[13]

First, the parameter *a* is the mean number of phonemes in monomorphemic words in a text. The parameter *b* determines the steepness of the curve, which corresponds to the rate of the shortening of the mean length of morphemes when words get longer (the smaller the value of *b*, the steeper the curve). Moreover, the values of *a* and *b* strongly correlate with the number of types in a text (the values of the Pearson correlation coefficient are 0.81 and -0.83, respectively), as well as with each other (0.93). These findings are consistent with the behaviour of the length of word types measured in syllables reported in [28] – the length of word types increases with the increasing text length. The positive correlation between the values of the parameter *a* and text length in word types indicates that the length of word types in morphemes behaves analogously. On the other hand, the negative correlation between the number of word types in a text and the value of parameter *b* suggests that the mean morpheme length decreases (with the increasing length of word types) more steeply in longer texts. The same interpretation of the parameters of the MAL at the level word – syllable – phoneme in Czech prosaic texts (and much weaker correlations in poems) can be found in [29].

A more precise and empirically based characterization of the interaction of the MAL with other language laws (such as, e.g., with the law of brevity, as discussed in Section 3) remains an open question for future research.

[13] The results, of course, depend on the segmentation rules we applied. However, they mostly follow the commonly accepted rules for the Czech language from [18] and [19], with the most important modification being a deeper segmentation for pronouns. As pronouns are a closed class of words, and we consider word types (and not word tokens), the change in their segmentation should not influence the results too much. Nevertheless, an analysis of the impact of segmentation rules (the choice of which is always at least partly subjective) can be an interesting topic for a future study.

ACKNOWLEDGEMENTS

The authors are thankful to Viktor Elšík (Department of Linguistics, Faculty of Arts, Charles University in Prague) for his advices on the morphological segmentation.

The research has been supported by grant "Podpora talentovaných studentů doktorského studia na Ostravské univerzitě III (ev. č. smlouvy 07359/2019/RRC)" (K. Pelegrinová) and by VEGA grant 2/0096/21 (J. Mačutek).

References

[1] Cramer, I. M. (2005). Das Menzerathsche Gesetz. In R. Köhler, G. Altmann and R. G. Piotrowski (eds.), Quantitative Linguistics. An International Handbook. Berlin/New York: de Gruyter, pages 659–688.

[2] Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B. D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., and Vidya, M. N. (2009). Word Frequency Studies. Berlin/New York: de Gruyter.

[3] Bentz, C., and Ferrer-i-Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In C. Bentz, G. Jäger and I. Yanovich (eds.), Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics. University of Tübingen, online publication system. Available at: https://publikationen.uni-tuebingen.de/xmlui/handle/10900/68558.

[4] Menzerath, P. (1954). Die Architektonik des deutschen Wortschatzes. Bonn: Dümmler.

[5] Altmann, G. (1980). Prolegomena to Menzerath's law. In R. Grothjahn (ed.), Glottometrika 2. Bochum: Brockmeyer, pages 1–10.

[6] Altmann, G., and Schwibbe, M. H. (1989). Das Menzerathsche Gesetz in informationsverarbeitenden Systemen. Hildesheim: Georg Olms Verlag.

[7] Kelih, E. (2010). Parameter interpretation of Menzerath law: Evidence from Serbian. In P. Grzybek, E. Kelih and J. Mačutek (eds.), Text and Language. Structures, Functions, Interrelations, Quantitative Perspectives. Wien: Praesens, pages 71–79.

[8] Mačutek, J., and Rovenchak, A. (2011). Canonical word forms: Menzerath-Altmann law, phonemic length and syllabic length. In E. Kelih, V. Levickij and V. Matskulyak (eds.), Issues in Quantitative Linguistics 2. Lüdenscheid: RAM-Verlag, pages 136–147.

[9] Mačutek, J., and Mikros, G. K. (2015). Menzerath-Altmann law for word length motifs. In G. K. Mikros and J. Mačutek (eds.), Sequences in Language and Text. Berlin/Boston: de Gruyter, pages 125–131.

[10] Teupenhayn, R., and Altmann, G. (1984). Clause length and Menzerath's law. In J. Boy and R. Köhler (eds.), Glottometrika 6. Bochum, Brockmeyer, pages 127–138.

[11] Mačutek, J., Čech, R., and Milička, J. (2017). Menzerath-Altmann law in syntactic dependency structure. In S. Montemagni and J. Nivre (eds.), Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017). Linköping: Linköping University Press, pages 100–107.

[12] Köhler, R. (2015). Linguistic motifs. In G. K. Mikros and J. Mačutek (eds.), Sequences in Language and Text. Berlin/Boston: de Gruyter, pages 89–108.

[13] Gerlach, R. (1982). Zur Überprüfung des Menzerath'schen Gesetzes im Bereich der Morphologie. In W. Lehfeldt and U. Strauss (eds.), Glottometrika 4. Bochum: Brockmeyer, pages 95–102.

[14] Krott, A. (1996). Some remarks on the relation between word length and morpheme length. Journal of Quantitative Linguistics, 3(1), pages 29–37.
[15] Hřebíček, L. (1995). Text Levels. Constructs, Constituents and the Menzerath-Altmann Law. Trier: WVT.
[16] Altmann, G., Erat, E., and Hřebíček, L. (1996). Word length distribution in Turkish texts. In P. Schmidt (ed.), Glottometrika 15. Trier: WVT, pages 195–204.
[17] Milička, J. (2014). Menzerath's law: The whole is greater than the sum of its parts. Journal of Quantitative Linguistics, 21(2), pages 85–99.
[18] Slavíčková, E. (1975). Retrográdní morfematický slovník češtiny s připojenými inventárními slovníky českých morfémů kořenových, prefixálních a sufixálních. Praha: Academia.
[19] Komárek, M., Kořenský, J., Petr, J., and Veselková, J. (1986). Mluvnice češtiny. Svazek 2. Tvarosloví. Praha: Academia.
[20] Komárek, M. (2016). Příspěvky k české morfologii. 2nd ed. Olomouc: Periplum.
[21] Altmann, G. (1992). Das Problem der Datenhomogenität. In B. Rieger (ed.), Glottometrika 13. Bochum: Brockmeyer, pages 287–298.
[22] Grzybek, P. (2013). Homogeneity and heterogeneity within language(s) and text(s): Theory and practice of word length modeling. In R. Köhler and G. Altmann (eds.), Issues in Quantitative Linguistics 3. Lüdenscheid: RAM-Verlag, pages 66–99.
[23] Williams, J. R., Bagrow, J. P., Danforth, C. M., and Dodds, P. S. (2015). Text mixing shapes the anatomy of rank-frequency distributions. Physical Review E, 91, 052811.
[24] Čech, R., Kosek, P., Mačutek, J., and Navrátilová, O. (2020). Proč (někdy) nemíchat texty aneb Text jako možná výchozí jednotka lingvistické analýzy. Naše řeč, 103(1–2), pages 24–36.
[25] Kubát, M. (2016). Kvantitativní analýza žánrů. Ostrava: Ostravská univerzita.
[26] Mačutek, J., Chromý, J., and Koščová, M. (2019). A data-based classification of Slavic languages: Indices of qualitative variation applied to grapheme frequencies. Journal of Quantitative Linguistics, 26(1), pages 66–80.
[27] Mačutek, J., and Wimmer, G. (2013). Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. Journal of Quantitative Linguistics, 20(3), pages 227–240.
[28] Kelih, E. (2012). On the dependency of word length on text length. Empirical results form Russian and Bulgarian parallel texts. In S. Naumann, P. Grzybek, R. Vulanović and G. Altmann (eds.), Synergetic Linguistics. Text and Language as Dynamic Systems. Wien: Praesens, pages 67–80.
[29] Čech, R., and Mačutek, J. (2021). The Menzerath-Altmann law in Czech poems by K. J. Erben. Proceedings of the Conference Plotting Poetry 4 (in print).

sciendo
DOI 10.2478/jazcas-2021-0038

PERSISTENT FEATURES – CORPUS-BASED EVIDENCE FOR REALLOCATION PROCESSES IN GERMAN

ELISABETH SCHERR
University of Graz, Graz, Austria

SCHERR, Elisabeth: Persistent features – Corpus-based evidence for reallocation processes in German. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 415 – 424.

Abstract: This study aims at tracing a reallocation process of a grammatical feature alongside the dialect-standard axis with the aid of corpus linguistics methods; more precisely with an integrative application of quantitative and qualitative approaches. The phenomenon under investigation is articles without the definiteness marker *d-* in German, usually ascribed to the Bavarian dialect area. Analyses show, however, that this apparently dialectal feature diffuses to other communication settings closer to the intended standard language use. This process is accompanied by a refunctionalisation of reduced article forms, indicating the relevance of language-internal relations for reallocation of grammatical features. The methodical approach should be easily applicable to other variants and – as many European languages show a diaglossic repertoire – relevant to other languages as well.

Keywords: reallocation, article system, Bavarian, dialect-standard axis

1 INTRODUCTION

In the vast majority of investigations dealing with the area of tension between a dialect and the standard[1], the data often speak for a general tendency of dialect reduction, levelling or loss, especially with regard to younger respondents [1]. Latest research on youth languages or on urban communication in German also seem to point in direction of a general (re-)standardisation tendency ([2], [3]). While the quantity of dialect features subjected to reductive change may be reason enough to assume a progressing limitation of dialect use, comparatively little attention has been paid to less frequent, yet nonetheless existing, persistent features.[2] Within the field of traditional dialectology, the difference between stability and change is associated with primary and secondary dialect features: Secondary features remain subconscious and are more stable, primary features are prone to change [5].

Whereas the approaches cited above discuss stability within one specific variety, few studies focus on stable features diffusing alongside the vertical dialect-

[1] *Dialect* and *standard language* are understood here as two (solely conceptual/theoretic) poles of a diaglossic continuum. This relation can be very specific with regard to the respective region (cf. the language situation in Switzerland or South Tyrol).

[2] Explaining the stability of forms which to a certain extent seem to be immune to linguistic change is addressed by Weinrich/Labov/Herzog [4] under the heading of the "actuation problem".

standard axis, with special focus on dialect features transferred to vertically higher[3] communication settings [6]. Expanding the traditional categorisation, such processes should affect solely "tertiary features" [7] that are usually very persistent and due to their low degree of salience[4] also easily transferable to more formal communication settings evoking intended standard language use.[5] This convergence may be accompanied by at least partly modified pragmatic, sociostylistic, or (as will be shown below) language-internal function of the original dialect features. With regard to dialect contact situations, Britain and Trudgill [10] call this process "reallocation". Up to date, the focus of reallocation studies mainly lies on phonological phenomena, whereas information on the behaviour of dialectal grammatical features diffusing closer to the intended standard is still a desideratum [11]. As a consequence, hypothesised reasons for such processes, elaborated on the basis of phonological features, have a limited explanatory potential with regard to grammar. In other words, the influence of social norms, identity building or unhindered communication ([12] and [7] amongst others) are arguably not so very well suited for explaining tertiary grammatical features, considering their mostly obligatory, subconscious use and low saliency.

2 APPROACHING REALLOCATION

Contrary to an extensive research on regional variation in general, empirical tracing of grammatical reallocation processes is still scarce, at least with regard to German. In-vogue methods in modern German dialectology like speech production tests or verbal and matched guise techniques that rest on deductive testing of an existing theory [13] are of limited use here, at least for two main reasons: (i) For a full, usage-based understanding of a potential grammatical reallocation phenomenon without considering a priori assumptions, a quantitative and qualitative corpus analysis of free speech production is obligatory. Besides the correlative-global data between the variable communication situation and the variant in question, semantic and pragmatic aspects of its conversational local use must be considered [14]. (ii) With regard to the corpus design, it is crucial that it reflects the horizontal (dialect) region(s), as well as the vertical (dialect-standard axis) dimension.

The following corpus study aims at empirically approaching the grammatical reallocation process. The linguistic features under investigation are the so-called "unstressed articles" ([15], [16]) mostly ascribed to Bavarian dialects. They are used

[3] The use of *higher* refers to the usual depiction of a vertical dialect-standard-continuum and does not indicate any other biased evaluation.

[4] For a discussion of the term *saliency* in sociolinguistics see [8].

[5] „[T]he situation of an interview with an unknown researcher is clearly one in which it is appropriate to use the standard." [9]

without the initial plosive *d-*, usually expressing definiteness of the associated noun, reducing the article to the derivational suffix (encoding number and case).

(1)	0580 DoG:	*die*	*miassen*	*letztendlich*	*∅**as***	*buach*	*lesen*
	Standard German:	die	müssen	letztendlich	**das**	buch	lesen
		they	need	in the end	the	book	read
		'they need to read the book in the end'					

The use of these reduced articles in dialectal speech seems to be highly frequent yet unsystematic[6], an observation which shell be scrutinized with regard to the vertical dialect-standard axis.

2.1 Data

The aim of the following corpus analysis is to trace potential diffusion processes alongside the dialect-standard axis in German with special focus on the dialect regions of Austria. These southern parts of the coherent German-speaking area are mostly characterised by a small-scale complex structure of dialects (cf. figure 1) and a dense diaglossic spectrum of variants[7] making for an ideal area of investigation regarding vertical reallocation processes. The data rest on a corpus of spoken language recorded in course of a Special Research Programme (SFB) "German in Austria. Variation – Contact – Perception"[8] (DiÖ). Project part 03 ("Speech repertoires and varietal spectra") focusses on rural areas of Austria, recording speakers in up to seven different settings. For the present study, the focus lies on data of 44 autochthonous speakers (20–30 years old[9]), each of them born in an Austrian village (13 research locations in total, see figure 1) and still living predominantly there or nearby. The probands were recorded in two different conversion settings to trace their intra-speaker variation spectrum: In a rather formal interview setting, the probands were prone to their (intended) standard language use, whereas a conversation among friends in absence of the interviewer triggered their most informal (dialectal) speech production. The specific corpus design is thus well suited for research questions touching upon the parameters of age, gender, degree of formality (correlated with closeness/distance to the intended standard language use) and region. For each of the dialect regions displayed in fig. 1, six to eight hours of speech recordings were analysed which makes for a total of 32 hours of interviews

[6] For a discussion of the phenomenon in Bavarian itself and regarding its unclear distribution cf. [15].

[7] The only exception from this general tendency is the westernmost part of Vorarlberg arguably showing a tendency toward a diglossic functional separation between the use of Alemannic dialect and standard German (see below).

[8] For an extensive discussion of the SFB cf. [17].

[9] For the relevance of a comparison with older speakers see chapter 3.

and 34 hours of conversations among friends. If not indicated otherwise, the following observations are limited to the analysis of the article inflection nominative singular neuter. The definite article *das* is the most frequent in the corpus and its analysis thus promises the highest degree of reliability.[10]

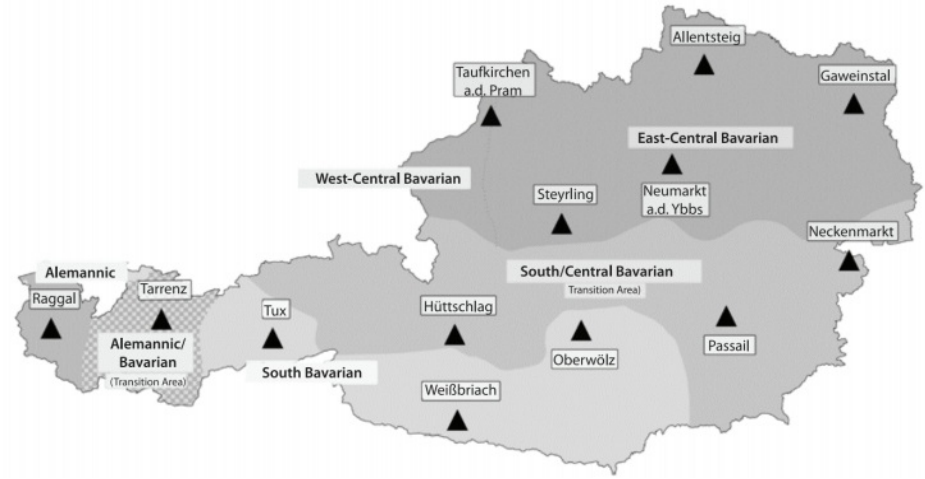


Fig. 1. Dialect areas in Austria and research locations (triangles) [19]

2.2 Preliminary results

The informal communication setting eliciting dialect use displays – as expected – very high relative frequencies of reduced article forms in nominal phrases in all of the analysed dialect regions. The ratios in the respective areas are surprisingly stable with a share of around 50% of reduced forms in most regions. Also, the Chi-squared test gives the p-value of 0.2256, which indicates that generally there is no significant difference in the data. The only outlier is the South/Central Bavarian transition area with a ratio of over 64% of reduced article forms.

	east-central		south-central		south		alemannic/ bavarian		alemannic	
	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.
reduced	115	50	147	**64**	96	51	18	54	35	50
non-reduced	114	50	83	36	93	49	39	46	35	50
∑	229		230		189		57		70	

Tab. 1. Absolute/relative frequencies of reduced/full article forms in the respective dialect regions in informal communication settings

[10] The comparatively high frequency of reduced *das*-forms certainly also has phonetical-phonological reasons, a phenomenon I cannot go further into here, see however [18].

Furthermore, the data seem to confirm the observations in previous studies ([18], [20]) that the deletion of the definiteness marker *d-* does not follow any systematic regularity: It appears in all contexts, also in those only full forms would be acceptable in standard language use, inter alia expressing situation-deictic reference (2), as well as demonstrative (3) or anaphoric reference (4):

(2) 0206 *nehma* ***s_nägste karterl*** *liawa* *ne?*
take we the next card rather not?
'let's rather take the next card, shell we?'

(3) 0471 *besonderes wort im dialekt (()) oachkatzlschwoaf*
'special word in [your] dialect (()) tail of a squirrel'
0491 *jo* ***es wort*** *is cool*
'yes this word is cool'

(4) 0027 ***is madl*** *wor* *holt* *imma* *so* *vul* *gestresst*
'this/the girl was just always so extremely stressed out'

The localisation of specific definite reference objects [21] should generally favour the use of full article forms in these contexts, thus dropping of the definiteness marker *d-* in contexts such as (2)–(4) contradicts the central function of definiteness. The seemingly chaotic use and the high frequency of reduced articles in dialect use have led authors to argue for a revocation of the German article system in general with advanced stages in some dialects indicating the progressing loss of definite articles ([18], [20]).

Looking at the interview setting, the quantitative analysis of these formal communications clearly shows a decline of reduced article forms with a ratio of around 30% in most dialect regions (see table 2). Two regions, however, stand out: The South/Central transition area, again with a higher ratio of 49%, and the Alemannic dialect region of (most parts of) Vorarlberg, with only 18% of all nominal phrases showing reduced articles. This last point suggests that the growing diglossic relations between dialect and standard in the westernmost parts of Austria arguably cause a shift to standard-close variants, rendering the communication comparatively less influenced by (Alemannic) dialect features.[11] Contrary to the data of the informal setting, these differences are highly significant with p=0.0001501.

	East-Central Bavarian		South/Central Bavarian		South Bavarian		Alemannic/ Bavarian		Alemannic	
	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.
reduced	53	31	81	**49**	62	30	9	31	24	**18**
non-reduced	116	69	86	51	143	70	20	69	109	82
∑	169		167		205		29		133	

Tab. 2. Absolute/relative frequencies of reduced/full article form in the respective dialect regions in formal communication settings

[11] For a discussion of the complex dialect-pragmatic status of Vorarlberg see [22].

Apart from the expected decline of dialect features in formal communication settings accompanied by less frequent use of the reduced articles, it is surprising that in most regions still 30% of all nominal phrases – in the South/Central transition area even half! – are used with reduced articles. To probe into this outcome, the quantitative results were complemented with a conversational local analysis focusing on the functional value of the reduced article forms in formal communication settings. Qualitative analyses strikingly show that articles without the definiteness marker *d-* mainly appear in formal communication settings if and only if they refer to an abstract, non-localizable or exclusive entity. Thus, they do not fulfil the core function of definiteness as they express what Ágel [23] calls "reine Aktualisierung" 'sheer activisation' of a concept without a deictic pointing relation. In other words: Contrary to the unsystematic use in dialect communication, in communication settings closer to the intended standard language, reduced articles are refunctionalised as means of expressing grammatical (inflectional) information without giving indication of localizing a limited, hence definite, entity. The examples (5)–(7) illustrate 96% of all cases without *d-* not containing any definite, let alone demonstrative, reference (see table 3). This functional preference explains their significantly frequent collocation with abstract nouns, unique nouns or nominalized adjectives as the latter are not prototypically associated with concrete localisation or limited reference.

	non-deictic	deictic/demonstrative
reduced	**96.06**	3.94
non-reduced	40.08	59.92

Tab. 3. Absolute/relative frequencies of reduced/full article form in the respective dialect regions in formal communication settings

(5) 024 ***is positive*** *an dem dialekt is*
the positive with the dialect is'
'the positive side of the dialect is…'

(6) 0245 *des is bei uns* ***as schifahrn***
this is for us skiing
'for us it is skiing'

(7) 0029 *i hob eben afoch net* ***is gfü*** *dass*
I have just simply not the feeling that
'I just do not have the feeling that…'

This clear tendency for reduced articles to appear with non-deictic, non-demonstrative referents is not to say, however, that this semantic group of nouns is never used with full article forms (see table 4) nor that concrete nouns never appear with reduced articles. To reveal their status as refunctionalized features, the following

analysis is limited to abstract nouns to shed light on their functional value in a specific context: Articles without the definiteness marker *d-* are predominantly used with reduced article forms in formal communication settings, rendering their ratios in every region significantly higher in comparison to table 2 (see table 4).

	East-Central Bavarian		South/Central Bavarian		South Bavarian		Alemannic/ Bavarian		Alemannic	
	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.
reduced	92	67	115	74	33	51	16	89	33	**67**
non-reduced	46	33	41	26	16	33	2	11	16	33

Tab. 4. Absolute/relative frequencies of reduced/full article form in the respective dialect regions in formal communication settings with restriction to abstract nouns

Particularly noticeable are the increased ratios in the Alemannic region (see also chapter 3). It shows the clearest picture with reduced article forms, generally being scarce in formal communication settings, their use with abstract nouns however is comparable to the other areas. The diagrams in figure 2 summarize the significant effects the limitation to functionally relevant contexts have on the evaluation of the status of reduced article forms as reallocation phenomena. With p-values ranging from p=0.0004874 (South/Central Bavarian) to p=7.393e-07 (East-Central Bavarian) the differences resulting from the limitation on abstract nouns prove to be highly significant in all regions.

3 CONCLUSION

The conducted study reveals substantial differences in the use of article forms without the definiteness marker *-d* in dialectal speech compared to communication settings closer to the intended standard language use. Research on this feature so far ascribes it a rather unsystematic, yet frequent use, which has led some authors to extrapolate the ongoing revocation of the article system. Instead of assuming such an erosion or future loss, a more detailed quantitative and qualitative analyses suggest that whereas reduced articles in fact seem to be used in dialectal communication rather unsystematically, they show a significant functional value in intended standard language use. Their surprisingly high frequency in formal communication settings, and especially their collocation with nouns referring to non-deictic/demonstrative referents, in fact speak for an ongoing reallocation process of a dialect feature accompanied with functional differentiation. It seems that newly developed functions of grammatical, "tertiary" [7] phenomena like the one discussed above rather touch upon inner-linguistic structural or functional relations than upon sociostylistic or allophonic reasons [10]: Grammatical vertical diffusion processes are arguably enhanced when they offer a systematic benefit.

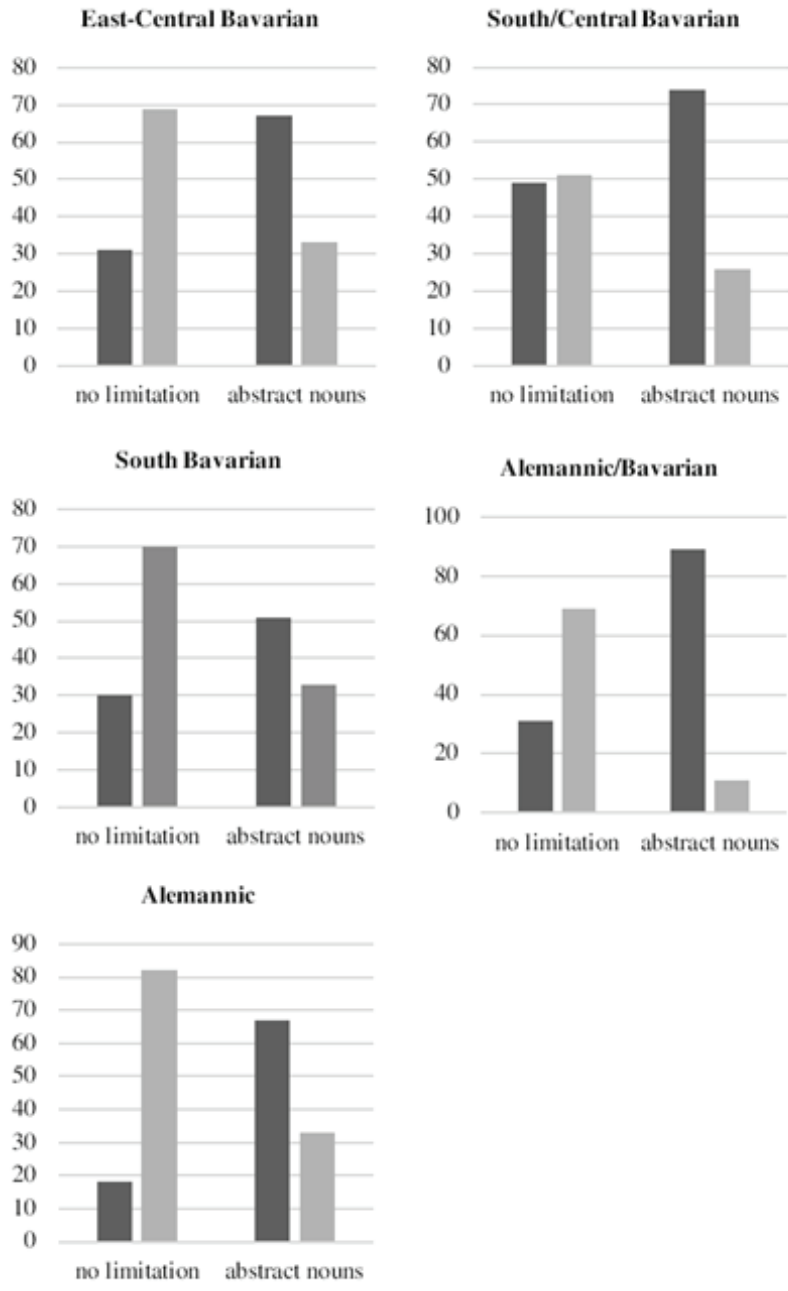


Fig. 2. Ratios of reduced (dark grey)/non-reduced (light grey) forms in all nominal phrases vs. with abstract nouns in all Austrian dialect regions

Finally, some qualifying point must be mentioned: Certainly, a range of follow-up-studies are necessary to confirm these observations with other inflectional forms of the article, with other semantic groups of nouns, with other dialect areas and – especially – with other age groups. With regard to that last point, a real- or apparent-time study could shed light on the question if the specific use of reduced articles in formal settings in fact indicates ongoing language change. Pilot studies regarding the age factor seem to confirm the status of reduced article form as reallocation phenomena; their collocation with unique nouns also underpins the findings with abstract nouns, the results of which I cannot demonstrate here for reasons of space. Nevertheless, the findings presented here have shown that it is not always the commonly assumed levelling and simplification processes of dialect features and their decline in favour of variants closer to the standard-pole that cause a change in the variant spectrum. An added functional value and low saliency of a (tertiary) dialect feature may pave the way for reallocation processes that are traceable with differentiated quantitative and qualitative corpus-linguistic methods.

ACKNOWLEDGEMENTS

Research for this contribution was conducted in the SFB "German in Austria. Variation – Contact – Perception" (Austrian Science Fund (FWF) 060, principal investigator: Alexandra N. Lenz).

References

[1] Rundblad, G. (1998). Stability in the theory of language change. In Papers from the 16th Scandinavian Conference of Linguistics, pages 369–380, Turku. Turunyliopisto.

[2] Ziegler, A. (2018). Undoing Youth – Dialect levelling and restandardisation in urban vernaculars in Austria. In Youth Languages. Current perspectives of international research, pages 49–65, Berlin and Boston. de Gruyter Mouton.

[3] Glauninger, M. (2010). Zwischen Hochdeutsch, Dialekt und Denglisch. Innere Mehrsprachigkeit und urbane Kommunikation am Beispiel der Jugendlichen im Ballungsraum Wien. In Übergang. Kommunikation in der Stadt und an ihren Rändern, pages 181–193, Linz. Adalbert-Stifter-Institut des Landes Oberösterreich.

[4] Weinreich, U., Labov, W., and Herzog, M. (1968). Empirical foundations for a theory of language change. In Directions for Historical Linguistics, pages 97–195, Austin/London. University of Texas Press.

[5] Schirmunski, V. (1930). Sprachgeschichte und Siedlungsmundarten. Germanisch-romanische Monatsschrift, 18, pages 113–122 and 171–188.

[6] Schwarz, C. (2020). Reduction and persistence of phonological dialect features in German. In Intermediate language varieties, pages 103–124, Amsterdam/Philadelphia. Benjamins.

[7] Taeldeman, J. (2009). Linguistic stability in a language space. In Language and Space. An International Handbook of Linguistic Variation. Volume 1: Theories and Methods, pages 355–374, Berlin. de Gruyter Mouton.

[8] Auer, P. (2004). Anmerkungen zum Salienzbegriff in der Soziolinguistik. Linguistik Online, 66(4), pages 7–20.
[9] Auer, P., and Spiekermann, H. (2011). Demotisation of the standard variety or destandardisation? In Standard Languages and Language Standards in a Changing Europe, pages 161–176, Oslo. Novus.
[10] Britain, D., and Trudgill, P. (2005). New Dialect Formation and Contact-induced Reallocation. International Journal of English Studies, 5(1), pages 183–209.
[11] Schmidt, J. et al. (2019). Sprache und Raum im Deutschen: Aktuelle Entwicklungen und Forschungsdesiderate. In Sprache und Raum. Ein internationales Handbuch der Sprachvariation, pages 28–60, Berlin and Boston. de Gruyter.
[12] Schmidt, J., and Herrgen, J. (2011). Sprachdynamik. Eine Einführung in die moderne Regionalsprachenforschung. Berlin: Schmidt, 464 p.
[13] Koppensteiner, W., and Lenz, A. (2020). Tracing a standard language in Austria using methodological microvariations of Verbal and Matched Guise Technique. Linguistic Online, 102(2), pages 47–82.
[14] Gilles, P. (2003). Zugänge zum Substandard. Korrelativ-globale und konversationell-lokale Verfahren. In Standardfragen. Soziolinguistische Perspektiven auf Sprachgeschichte, Sprachkontakt und Sprachvariation, pages 195–215, Frankfurt am Main et al. Lang.
[15] Weiß, H. (1998). Syntax des Bairischen. Studien zur Grammatik einer natürlichen Sprache. Tübingen: Niemeyer, 304 p.
[16] Zehetner, L. (1985). Das bairische Dialektbuch. München: Beck, 301 p.
[17] Lenz, A. (2018). The Special Research Programme German in Austria. Variation – Contact – Perception. In Language Choice in Tourism – Focus on Europe, pages 269–277, Berlin and Boston. de Gruyter.
[18] Nübling, D. (2005). Von in die über in'n und ins bis im. Die Klitisierung von Präposition und Artikel als ‚Grammatikalisierungsbaustelle'. In Grammatikalisierung im Deutschen, pages 105–132, Berlin and New York. de Gruyter.
[19] Fingerhuth, M., and Lenz, A. (2020). Variation and dynamics of complementizer agreement in German. Linguistic Variation 21(2), pages 1–48.
[20] Leiss, E. (2010). Artikel und Aspekt. Die grammatischen Muster von Definitheit. Berlin and New York: de Gruyter, 317 p.
[21] Lyons, C. (1999). Definiteness. Cambridge: University Press, 402 p.
[22] Ender, A., and Kaiser, I. (2014). Diglossie oder Standard-Dialekt-Standard-Kontinuum? Zwischen kollektiver, individueller, wahrgenommener und tatsächlicher Sprachvariation in Vorarlberg und im bairischsprachigen Österreich. In Alemannische Dialektologie. Dialekte im Kontakt, pages 131–146, Stuttgart: Steiner.
[23] Ágel, V. (1996). Finites Substantiv. Zeitschrift für Germanistische Linguistik 24, pages 16–57.

sciendo
DOI 10.2478/jazcas-2021-0039

ON CORPUS-DRIVEN RESEARCH OF COMPLEX ADVERBIAL PREPOSITIONS WITH SPATIAL MEANING IN CZECH

AKSANA SCHILLOVÁ
Czech Language Institute, Czech Academy of Sciences, Prague, Czech Republic

SCHILLOVÁ, Aksana: On corpus-driven research of complex adverbial prepositions with spatial meaning in Czech. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 425 – 433

Abstract: Complex adverbial prepositions with spatial meaning have not been sufficiently studied so far in Czech. To establish a set of these expressions in their actual usage, the resources of the Czech National Corpus were used in this study. The research has shown that the SYN2020 corpus is a relevant tool for searching for two-word expressions with a LOCATIVE ADVERB – SIMPLE PREPOSITION structure that have the same function as a one-word locative preposition. The article describes a method for the extraction of these expressions from the corpus, as well as a method for the collection of their quantitative data using corpus tools. As a result of the research, a list of expressions that are presumably complex prepositions is provided.

Keywords: complex preposition, locative adverb, spatial meaning, Czech language, Czech National Corpus

1 INTRODUCTION

The article deals with the use of the Czech National Corpus (CNC) in the study of two-word expressions that have a LOCATIVE ADVERB – SIMPLE PREPOSITION structure and can function as a complex preposition with spatial meaning. In Czech linguistic literature, only four adverbial complex prepositions with spatial meaning are described, or at least mentioned: *stranou od* 'aside from' ([1], [2], [3], [4], [5]), *daleko od* 'far from' [3]; *napravo od* 'to the right of' and *nalevo od* 'to the left of' [1]. Since a number of these units are detected and described in other Slavic languages (see the next section), it can be assumed that these complex prepositions also occur in Czech but have not been studied yet. Cf. the explanatory dictionary of Belarusian prepositions [6] that presents a list of "adverbial-prepositional constructions that function as prepositions" consisting of 127 units, 94 of which have a spatial meaning [6, pp. 165–166].

Since this type of complex prepositions in Czech has not been sufficiently covered in linguistic literature, the study was focused on the search for these complex units in actual language use, namely in a language corpus. The search was carried out on the SYN2020 corpus that is a 100m representative corpus of contemporary written Czech available within the CNC project [7].

This paper describes the extraction of statistically significant ADVERB – PREPOSITION combinations from the corpus using such corpus tools as advanced queries, positive/negative filters, frequency distribution, collocation candidates, association measures (T-score, MI-score, logDice), etc. The extracted combinations are non-random from statistical point of view and can be considered candidates for prepositionalization due to their fixity, idiomaticity, and frequency.

The expressions with the ADVERB – PREPOSITION structure have been already discussed in a 1977 article by Kroupova [8] as having the potential for prepositionalization. At the present day, we can evaluate these expressions in the light of corpus data.

2 THEORETICAL BACKGROUND

The present paper is mainly inspired by the following motivations:

1) Czech adverbs, as well as complex prepositions, are insufficiently covered in linguistic literature.

As for adverbs, only a few pages are devoted to their general description in Czech grammars (see [3], [9], [10], [11]). Moreover, the grammars focus mainly on description of adverb formation and degrees of comparison of adverbs. However, for example, the valence properties of adverbs have not been studied in detail [12], neither the limits of this word class, and the recognition criteria for adverbialization, too, remain terra incognita ([13], [14], [15]).

As for Czech complex prepositions, the monographs [1] and [5] make a significant contribution to their description. Nevertheless, these works primarily focus on the description of prepositional expressions that are derived from nouns, cf. the lists of prepositions represented in [1, pp. 323–333] and in [5, pp. 39–49]. The Czech prepositional system, taken as a whole, does not have yet a comprehensive linguistic description at the level of a monograph. Besides, there is no dictionary of Czech prepositions and prepositional expressions, while there are, for example, dictionaries of Ukrainian, Belarusian, Russian, Polish prepositions, and their analogues ([6], [16], [17], [18], [19]).

2) The study of complex adverbial prepositions, which are a neglected area of Czech grammar, can contribute to the lexical database of the LEMUR project [20].

The main task of this project is to create a new electronic linguistic resource, namely a database of Czech multiword expressions, which will subsequently be useful for many reasons, e.g., for teaching Czech as a foreign language, for lexicography, for the creation and improvement of natural language automatic processing tools ([20], [21], [22]). In this database, Czech multiword expressions will be presented and comprehensively described. The project also develops a typology of these units, in which complex prepositions are considered one of the syntactic types of Czech multiword expressions [22, p. 44].

3 DATA EXTRACTION FROM THE SYN2020 CORPUS

3.1 Methods of prepositionalization candidates extraction from the corpus

When we use representative SYN-series corpora [23], we must remember that these corpora contain not only original Czech texts but also texts translated into Czech from other languages.[1] Thus, to exclude the influence of a source language on the results of the present study, a subcorpus within the SYN2020 corpus was created consisting only of original Czech texts (size in tokens ca. 80m).

In this subcorpus, the search was carried out through the following advanced query: [tag="D.*"][tag="R.*"]. It means that all the possible ADVERB – PREPOSITION combinations which occur in the corpus were searched.[2] As a result, a concordance spanning 474.254 hits was generated.

The next step was to obtain a frequency list of combinations that were found. To make the frequency list, the menu item Frequency (Frequency Custom… > Frequency distribution) was used. The resulting list consisted of more than 20 thousand combinations arranged in descending order of their absolute frequency.

Since there is no semantic annotation in the SYN corpora, combinations with spatial meaning were selected manually from the list. The selection was made from the first thousand most frequent combinations. In a future study, this sample can be expanded to include less frequent combinations that have a rank higher than 1000.

3.2 Methods of quantitative data collection

After extracting data on the absolute frequency of the ADVERB – PREPOSITION combinations (see above), their relative frequency was calculated, i.e., a correlation between the absolute frequency of the entire combination and that of its adverb.[3] A relative frequency that is above average can be considered as a sign of stability of a given combination, cf. [1, p. 50].

In addition to the data on the absolute and relative frequency of the expressions studied in this paper, their statistical values were also analysed. For this purpose, using the menu item Collocations, collocation lists for the locative adverbs that are members of these expressions were obtained. The span of collocations was restricted to the first position to the right from a key word (a locative adverb). From the

[1] On the language of translations into Czech (see [24], [25], [26]); on the specialized JEROME linguistic corpus for analysing translated Czech, see [27].

[2] Previously, a similar search was carried out for denominal complex prepositions on the SYN2000 corpus by Blatná [1, p. 11]. Cf., the use of the p-collocation tool to search for Czech verbal participles as candidates for prepositionalization by Richterová [28].

[3] In other words, for each combination under study, it was checked how many times the adverb has occurred in the corpus and how many times the corresponding ADVERB – PREPOSITION combination has occurred in the corpus. Then, the percentage was calculated using a simple mathematical formula: Rel. freq. = A × 100 ÷ B, where A is the absolute frequency of the combination, B is the absolute frequency of the adverb that is a component of this combination.

collocation lists, simple prepositions were selected and data on the association measures (MI, T-score, logDice)[4] of the LOCATIVE ADVERB – SIMPLE PREPOSITION collocations were extracted.

Association measures help to identify which co-occurrences of words in a corpus are regular and non-random, namely from the point of view of statistics. This means, in terms of the ratio of the number of occurrences of collocations to the number of occurrences of their components taken separately and to the total number of all words in a corpus, see [29, pp. 103–105].

In the present study, it is assumed that the higher the measures of the statistical association between a locative adverb and a simple preposition, the more likely it is that the collocation is not a free combination of words but a fixed multiword expression that has the same function as a simple preposition.

At the next stage, the data of all the analysed combinations were compared relative to each other. As a result, it was revealed which of the combinations are statistically significant and hence can be considered potential complex prepositions. According to Petkevic et al. [22], complex prepositions show the so-called statistical idiomaticity, which means that these expressions are usually not semantically idiomatic but have an above-average frequency and a restricted collocability [22, p. 52].

4 RESULTS

In the subcorpus of original Czech texts (size in tokens ca. 80m), which was created within the SYN2020 corpus, 20.072 ADVERB – PREPOSITION combinations were found.

In the top ten most frequent combinations, there are two expressions that are already considered as complex prepositions in linguistic literature. These are *spolu s* ‘together with’ (rank 2) and *společně s* ‘jointly with’ (rank 7).[5] See Fig. 1.

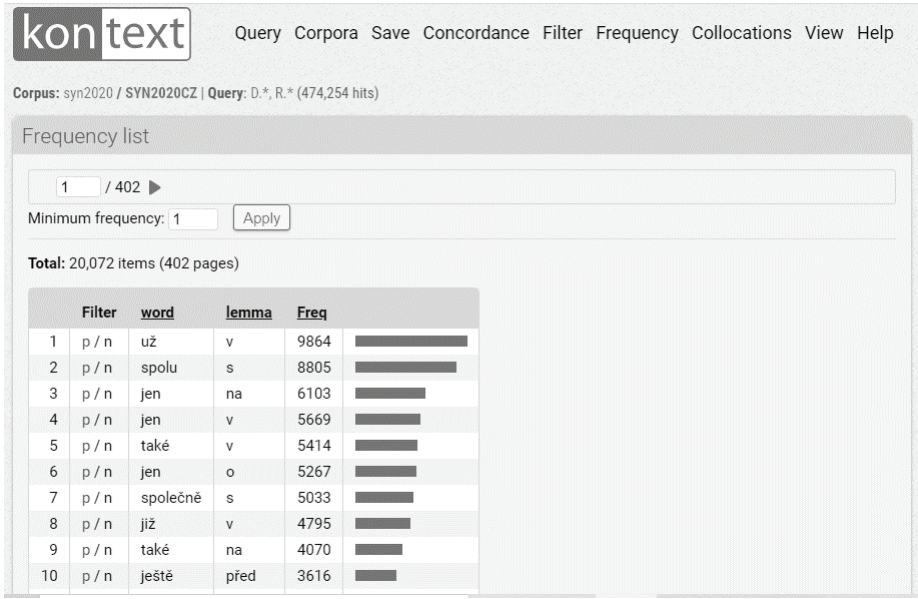
From the first thousand items of this frequency list, combinations with locative adverbs were manually selected: 61 expressions in total. The most frequent expressions are the following ones (their absolute frequency is given in brackets): *daleko od* ‘far from’ (779), *blízko k* ‘close to’ (371), *hluboko do* ‘deep into’ (353), *vysoko nad* ‘high above’ (317), *severně od* ‘north of’ (317), *daleko za* ‘far behind’ (298), *jižně od* ‘south of’ (292).

Some expressions are clearly distinguished from the other ones in terms of their relative frequency. This is a group of expressions that are used to refer to a location according to the cardinal direction (north, east, south, or west). These are the following ones: *severovýchodně od* ‘northeast of’, *severozápadně od* ‘northwest of’,

[4] For what these statistical measures mean and how they are calculated, see the corpus wiki accessible at: https://wiki.korpus.cz/doku.php/pojmy:asociacni_miry.

[5] The expression *spolu s* ‘together with’ is considered as a complex preposition in [1], [2], [3], [8], [9], [10], [30]; *společně s* ‘jointly with’ – in [1], [2], [5], [30].

jihovýchodně od ‘southeast of’, *jihozápadně od* ‘southwest of’, *východně od* ‘east of’, *severně od* ‘north of’, *jižně od* ‘south of’, *západně od* ‘west of’, which have a relative frequency of 82 to 92 %.



	Filter	word	lemma	Freq
1	p / n	už	v	9864
2	p / n	spolu	s	8805
3	p / n	jen	na	6103
4	p / n	jen	v	5669
5	p / n	také	v	5414
6	p / n	jen	o	5267
7	p / n	společně	s	5033
8	p / n	již	v	4795
9	p / n	také	na	4070
10	p / n	ještě	před	3616

Fig. 1. The beginning of the frequency list of the ADVERB – PREPOSITION combinations

For example, the adverb *severovýchodně* ‘northeast’ occurs in the subcorpus 86 times in 48 texts. However, in 79 cases, it co-occurs with the preposition *od* ‘of’, as in example (1) below, which is 92 % of the total number of the occurrences of the adverb in the subcorpus.

As for collocates on the right, the combination *severovýchodně od* ‘northeast of’ co-occurs exclusively (a) with toponyms, e.g., *Brno* ‘Brno’, *Bratislava* ‘Bratislava’, *Krakov* ‘Krakow’, *Japonsko* ‘Japan’, *Beringovy užiny* ‘Bering straits’, etc., 67 % of all the cases, or (b) with nouns referring to a place or an object in space, e.g., *obec* ‘municipality’, *vesnice* ‘village’, *město* ‘town’, *kostel* ‘church’, *nádraží* ‘station’, etc., 33 % of all the cases.

As for collocates on the left, the combination *severovýchodně od* ‘northeast of’ does not show the restricted collocability. In addition, it can be used as an element of an adverbial modifier separated by commas, as in example (2) below.

(1) *Farma leží* ***severovýchodně od Pekingu*** *a je největším zařízením svého druhu v Asii.* ‘The farm is located northeast of Beijing and is the largest facility of its kind in Asia.’

(2) *Na polích kolem Ovčár,* ***severovýchodně od Kolína****, můžeme i dnes najít valouny chalcedonu, jaspisu i křemene, pocházejících rovněž z labských teras.*
'In the fields around Ovcary, northeast of Kolin, we can still find boulders of chalcedony, jasper and quartz, which also come from the Elbe terraces.'

Moreover, the combination *severovýchodně od* 'northeast of' has a high association measure MI-score: 9.402. The other combinations of this semantic group (*severozápadně od* 'northwest of', *jihovýchodně od* 'southeast of', *jihozápadně od* 'southwest of', *východně od* 'east of', *severně od* 'north of', *jižně od* 'south of', *západně od* 'west of') also have MI > 7. Note that the boundary MI = 7 is considered relevant for systemic collocations [31]. Thus, from statistical point of view, these word combinations are systemic, fixed, and non-random.

The statistical data of all the combinations were summarized in one table and sorted by MI (from the highest to the lowest value). A fragment of the table (its beginning) that includes combinations with MI > 7 is presented in Table 1, where *Rank* is the position of the combination in the frequency list of bigrams [tag="D.*"] [tag="R.*"] sorted by their absolute frequency in descending order; *Abs. freq.* is the absolute frequency of the combination in the subcorpus, *Rel. freq.* is the percentage ratio of the absolute frequency of the combination to the absolute frequency of the adverb that is the left component of the combination.

Rank	**Combination**	**Abs. freq.**	**Rel. freq., %**	**MI ↓**	**T-score**	**logDice**
976	*nízko nad* 'low above'	75	14,8	10.743	8.655	5.810
259	*vysoko nad* 'high above'	317	18,5	10.445	17.792	7.873
422	*východně od* 'east of'	190	89,2	9.415	13.764	5.146
938	*severovýchodně od* 'northeast of'	79	91,9	9.402	8.875	3.881
285	*jižně od* 'south of'	292	83,9	9.389	17.063	5.765
260	*severně od* 'north of'	317	86,6	9.384	17.778	5.883
419	*západně od* 'west of'	192	82,4	9.379	13.836	5.161
953	*severozápadně od* 'northwest of'	77	90,6	9.365	8.762	3.844
305	*hluboko pod* 'deep below'	274	18,5	9.346	16.528	7.425
877	*jihovýchodně od* 'southeast of'	84	89,4	9.276	9.150	3.969

Rank	Combination	Abs. freq.	Rel. freq., %	MI ↓	T-score	logDice
981	*jihozápadně od* ‘southwest of’	75	89,3	9.274	8.646	3.806
827	*napravo od* ‘right of’	90	22,9	9.202	9.471	4.069
597	*nedaleko od* ‘not far from’	132	23,3	8.975	11.466	4.621
84	*daleko od* ‘far from’	779	8,9	8.328	27.824	7.169
856	*dole pod* ‘down below’	87	3,2	8.268	9.297	5.779
766	*vpravo od* ‘right of’	97	3,6	7.961	9.809	4.175
213	*blízko k* ‘near to’	371	12,1	7.896	19.181	5.132
978	*vlevo od* ‘left of’	75	2,5	7.487	8.612	3.804

Tab. 1. Quantitative data of the LOCATIVE ADVERB – SIMPLE PREPOSITION combinations that have MI > 7

In a future study, the combinations as used in the corpus will be analysed in terms of their semantics, collocability, and syntactic behaviour. The qualitative analysis will help to establish (a) which of them are actually used as complex prepositions, (b) under what conditions they have a prepositional function, (c) at what stage of prepositionalization they currently are.

5 CONCLUSION

The present study has shown that the search and statistical tools of the SYN2020 corpus are appropriate to detect the ADVERB – PREPOSITION expressions that are presumably used as complex prepositions with spatial meaning. The quantitative data extracted from the corpus serve primarily as an indicator of the fixity, regularity, and non-randomness of these expressions, which allows them to be detected.

Nevertheless, it should be noted that the quantitative data alone are not sufficient to claim that the expressions investigated in the present paper are prepositions. For this purpose, it is necessary to develop a special methodology that will be based on a close qualitative analysis of their actual usage. There are already some developments for recognizing the prepositional function of Czech denominal expressions, see [1] and [5]. However, the specific features of an adverb as a part of speech (the indeclinability, the heterogeneity of this word-class in terms of origin, semantics, syntactic functions, etc.) require specific analysis methods, which have

not yet been developed for the Czech language. The development of such methodology will be a task for further research.

References

[1] Blatná, R. (2006). Víceslovné předložky v současné češtině. Praha: NLN, Nakladatelství Lidové noviny, 351 p.

[2] F. Čermák, J. Hronek and J. Machač (eds.). (1988). Slovník české frazeologie a idiomatiky: Výrazy neslovesné. Praha: Academia, 511 p.

[3] Karlík, P., Nekula, M., Rusínová, Z., and Grepl, M. (2012). Příruční mluvnice češtiny. Praha: NLN, Nakladatelství Lidové noviny, 799 p.

[4] M. Komárek, J. Kořenský, J. Petr and J. Veselková (eds.). (1986). Mluvnice češtiny 2: Tvarosloví. Praha: Academia, 536 p.

[5] Kroupová, L. (1985). Sekundární předložky v současné spisovné češtině. Praha: Ústav pro jazyk český ČSAV, 155 p.

[6] Šuba, P. (1993). Tlumačalny slounik belaruskich prynazounikau. Minsk: Narodnaja asveta, 168 p.

[7] Křen, M., Cvrček, V., Henyš, J., Hnátková, M., Jelínek, T., Kocek, J., Kováříková, D., Křivan, J., Milička, J., Petkevič, V., Procházka, P., Skoumalová, H., Šindlerová, J., and Škrabal, M. (2020). SYN2020: reprezentativní korpus psané češtiny. Ústav Českého národního korpusu FF UK, Praha. Accessible at: http://www.korpus.cz.

[8] Kroupová, L. (1977). Další sporné případy sekundárních předložek. In Naše řeč, 60(2), pages 68–75.

[9] Čechová, M., Dokulil, M., Hlavsa, Z., Hrbáček, J., and Hrušková, Z. (2011). Čeština – řeč a jazyk. Praha: SPN – pedagogické nakladatelství, 442 p.

[10] Štícha, F. et al. (2018). Velká akademická gramatika současné češtiny. I(1). Praha: Academia, 763 p.

[11] Štícha, F. et al. (2013). Akademická gramatika spisovné češtiny. Praha: Academia, 974 p.

[12] Sláma, J., and Štěpánková, B. (2019). On the Valency of Various Types of Adverbs and Its Lexicographic Description. In Jazykovedný Časopis (Philological Journal) [Online], 70(2), pages 158–169.

[13] Vondráček M. (2020). Výrazy typu *běda* z pohledu slovnědruhového. In Lingvistika – Korpus – Empirie. Praha: Ústav pro jazyk český, pages 93–102.

[14] Vondráček, M. (2018). Slovnědruhové přechody, sporná slovnědruhová klasifikace. In F. Štícha et al., Velká akademická gramatika současné češtiny. I(1). Praha: Academia, pages 100–107.

[15] Vondráček, M. (1999). Příslovce a částice – hranice slovního druhu. In Naše řeč, 82(2), pages 72–78.

[16] Zagnitko, A., Daniluk, I., Sitar, G., and Sukina, I. (2007). Slovnik ukrajinskich prijmennikiv. Sučasna ukrajinska mova. Doneck: TOV VKF "BAO", 416 p.

[17] Kanjuškevič, M. (2008). Belaruskija prynazouniki i ich analahi. Hramatyka realnaha užyvannja. Materyjaly da slounika. Hrodna: HrDU, 492 p.

[18] Vsevolodova, M., Vinogradova, E., and Čaplygina, T. (2018). Russkie predlogi i sredstva predložnogo tipa. Materialy k funkcionalno-grammatičeskomu opisaniju realnogo upotreblenija. Moskva: URSS, 800 p.

[19] Lachur, Cz. (2019). Polskie przyimki wtórne i jednostki o funkcji przyimkowej w użyciu realnym. Materiały do słownika (w zestawieniu z językiem rossijskim). Tom 1. Kępa, 425 p.
[20] Hnátková, M., Jelínek, T., Kopřivová, M., Petkevič, V., Rosen, A., Skoumalová, H., and Vondřička, P. (2019). Lexical database of multiword expressions in Czech. In V. Zakharov (ed.), Trudy meždunarodnoj konferencii "Korpusnaja lingvistika – 2019", St. Petersburg: Saint Petersburg University Press, pages 9–16.
[21] Hnátková, M., Jelínek, T., Kopřivová, M., Petkevič, V., Rosen, A., Skoumalová, H., and Vondřička, P. (2018). Lepší vrabec v hrsti nežli holub na střeše. Víceslovné lexikální jednotky v češtině: typologie a slovník. Korpus – gramatika – axiologie, 9(17), pages 3–22.
[22] Petkevič, V., Kopřivová, M., Hnátková, M., Jelínek, T., Kopřiva, P., Rosen, A., Skoumalová, H., and Vondřička P. (2020). Typologie víceslovných jednotek v češtině a frekvenční zastoupení jejich hlavních vlastností v žánrově vyváženém korpusu. In Studie z aplikované lingvistiky/Studies in Applied Linguistics, 11, pages 37–62.
[23] Hnátková, M., Křen, M., Procházka, P., and Skoumalová, H. (2014). The SYN-series corpora of written Czech. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavík: ELRA, pages 160–164.
[24] A. Čermáková, L. Chlumská and M. Malá (eds.). (2016). Jazykové paralely. Praha: Nakladatelství Lidové noviny, 290 p.
[25] Chlumská, L. (2017). Překladová čeština a její charakteristiky. Praha: Nakladatelství Lidové noviny, 149 p.
[26] Chlumská, L., and Richterová, O. (2014). Překladová čeština v korpusech. In Naše řeč, 97(4–5), pages 259–269.
[27] Chlumská, L. (2013). JEROME: jednojazyčný srovnatelný korpus pro výzkum překladové češtiny. Ústav Českého národního korpusu FF UK, Praha. Accessible at: http://www.korpus.cz.
[28] Richterová, O. (2016). Identifikace posunů ve slovnědruhové příslušnosti: nejen na paralelních korpusech. In A. Čermáková, L. Chlumská and M. Malá (eds.), Jazykové paralely. Praha: Nakladatelství Lidové noviny, pages 95–144.
[29] Čermák, F. (2017). Korpus a korpusová lingvistika. Praha: Univerzita Karlova, nakladatelství Karolinum, 268 p.
[30] Cvrček, V. et al. (2015). Mluvnice současné češtiny 1. Jak se píše a jak se mluví. Praha: Univerzita Karlova v Praze, nakladatelství Karolinum, 416 p.
[31] J. Kocek, M. Kopřivová and K. Kučera (eds.). (2000). Český národní korpus: Úvod a příručka uživatele. Praha: FF UK v Praze, 156 p.

sciendo
DOI 10.2478/jazcas-2021-0040

THE STUDY OF VALENCY IS BIASED TOWARD MORE FREQUENT VERBS: A CORPUS STUDY OF THE VALENCY OF LESS FREQUENT VERBS IN CZECH

JAKUB SLÁMA
Czech Language Institute, Czech Academy of Sciences, Prague, Czech Republic

SLÁMA, Jakub: The study of valency is biased toward more frequent verbs: A corpus study of the valency of less frequent verbs in Czech. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 434 – 443.

Abstract: Theories of valency and valency dictionaries are inevitably and understandably based on the valency behavior of frequent verbs. This paper scrutinizes 154 low-frequency Czech verbs and argues that they demonstrate that Czech verbs are more malleable in their valency behavior than suggested by the literature. It is argued that this fits better within a constructionist approach to valency rather than a lexicalist one. Furthermore, the paper illustrates two alternations, previously unrecognized for Czech as semantic diatheses, namely the causative-inchoative alternation and the Agent-Means alternation.
Keywords: valency, valency alternation, causativity, frequency

1 INTRODUCTION[1]

What we (think we) know about valency (in Czech) is somewhat biased toward more frequent verbs. Valency theories are based on examples featuring frequent predicates, and valency dictionaries understandably describe the valency behavior of the most frequent verbs (or words of other parts of speech, which are not the focus here, however). If we examine the behavior of less frequent verbs, we might encounter phenomena which might present difficulties for the traditional approaches to valency; consider, e.g., the following examples of the metaphorical sense of the infrequent verb *hypertrofovat*, roughly corresponding to 'grow':

(1) *Tento trend **hypertrofuje** zejména v posledních deseti letech.* (syn v8 [1])
'This trend has been growing especially in the last ten years.'
(2) *Komplikují a **hypertrofují** legislativu.* (syn v8)
'They are complicating the legislation and making it (grow) too complex.'
(3) *Ten **hypertrofoval** v podobu, kterou nelze finančně udržet.* (syn v8)
'It grew into a form that is impossible to sustain financially.'

[1] I would like to thank Václava Kettnerová for her comments concerning the valency frames and alternations discussed in the paper.

(4) *Tuhle svoji dětskou touhu jsem **hypertrofoval** do téhle chalupy.* (syn v8)
'I transformed this childhood dream of mine into this cottage.'

The traditional approach is to say that since we are dealing with multiple (presumably four) different valency frames, we are dealing with multiple senses of the verb. However, this is a prime example of what has been described as the polysemy fallacy, that is, of viewing contextually-bound uses of a lexical item as instances of polysemy [2, p. 63]. Even more problematically, the reasoning is cyclical [3, p. 10], as it would have us believe that we are dealing with four different valency frames precisely because we have four different senses of the verb, while deducing the four senses of the verb on the basis of its use in four different valency frames. Furthermore, the idea that using the same verb in different valency frames (as, e.g., in the locative alternation) leads to polysemy has been challenged by psycholinguistic evidence [4].

2 THEORETICAL BACKGROUND

2.1 Two basic approaches to valency

There are two types of approaches to valency, or, argument structure [5, pp. 11–12]. The traditional, lexicalist, approach is characterized by the central belief that the argument structure of a predicate is determined by the predicate itself and by its semantics, and that if a verb occurs in various valency frames, these are associated with various senses of the verb. The traditional approaches to valency known in the Czech context – that of Daneš and Hlavsa [6], and the Functional Generative Description [7] – are lexicalist in nature.

An alternative approach originated within the framework of Construction Grammar (for a brief introduction, see [8]), based on the observation that argument structure cannot, in fact, be trivially reduced to knowledge tied to individual verbs (or other predicates), as illustrated by examples such as the following, cited after [9, p. 2]:

(5) *He **stared** her into immobility.*
(6) *Chess **coughed** smoke out of his lungs.*
(7) *Her nose was so bloodied that the ref **whistled** her off the floor.*
(8) *Navin **sneezed** blue pollen onto his shirt.*

These examples feature an unusual use of the verb in bold; in light of such examples, "the idea that argument structure is primarily knowledge about verbs loses some of its appeal" [9, p. 2]. The basic idea of the constructionist approach is that argument structure constructions (ASCs) – which are constructions in the sense of Construction Grammar, that is, they are Saussurean signs – exist independently of

verbs and have their own abstract meaning, and verbs might be combined with various ASCs. In English, a common verb like kick might be used in as many as some nine ASCs, without any apparent changes in its semantics [10, p. 394]. The constructionist approach has received a lot of experimental support (reviewed in [11]), and a fundamentally similar view of valency is espoused in various approaches other than Construction Grammar, which independently arrive at the conclusion that valency patterns (or, ASCs) exist as autonomous Saussurean signs (e.g., [12], [13], [14], [15]).

2.2 Bias toward more frequent verbs

Quite understandably, Czech valency dictionaries such as *Vallex* [16] include the most frequent verbs of Czech, with the latest version including 4,659 of them, which is roughly 22% of all verbs found in the representative corpus of written Czech syn2020 [17]. Nevertheless, it might be interesting to inspect less frequent verbs with respect to valency, since this can allow us to scrutinize the valency behavior of a verb in its entirety, without necessarily limiting our scope of attention to a sample of its uses and to its most typical uses, which is inevitable when compiling a (valency) dictionary. Furthermore, it has been known for a long time that more frequent words are more prone to polysemy [18, p. 109], and so inspecting the use of low-frequency verbs might allow us to study their valency behavior without the burden of polysemy. Finally, especially within the framework of usage-based (cognitive) linguistics, it has been abundantly demonstrated that frequency plays a crucial role in language, and more frequent units or expressions might behave quite differently from less frequent ones (cf. [19]). The intuition that frequency is relevant has also been present in valency research (e.g., [13, p. 59]), although rather marginally (but cf. e.g. [20]).

3 DATA

From the corpus syn2020 [17] I extracted the frequency list of all verbs, from which I selected the 118 verbs that occur in the corpus twenty times and the 36 verbs that occur seventy times (both of these numbers are arbitrary). All 4,880 occurrences of the 154 verbs were manually inspected, 36 of them were discarded (mostly as errors in lemmatization), and the 154 verbs were annotated for their valency behavior, especially for the number and type of valency frames in which they occurred in the data. Unless otherwise specified, all the examples cited in the paper are from syn2020. Occasionally, I use handier examples found in the bigger corpus syn v8 [1].

In describing valency frames, I mostly followed the Functional Generative Description. When a verb was found in multiple grammatical diatheses such as the passive (cf. [21]), these were naturally not taken as constituting different valency

frames of the verb. The same applies to definite (*nyní spolurozhodují obyvatelé Vranovic* 'now the residents of Vranovice are codeciding') and indefinite null objects (*právo spolurozhodovat* 'the right to codecide'); on both see, e.g., [22]. Similarly, I did not take examples of reflexive objects to constitute a new valency frame; that is, the same verb with a single valency frame is instantiated in *ověsí se šperky* 'she will decorate herself with jewels' and *ověsí svoji polovičku blýskavějšími diamanty* 'he will decorate his partner with more glittering diamonds'.

On the other hand, I took instances such as the following, in which the reflexive variant does not denote an action whose patient is expressed by the reflexive *se*, to represent two different valency frames of the same verb. In this respect, I diverge from most traditional accounts, which would see the verbs as two different lexical units, and I do so simply because I view examples such as these two as representing an identical meaning of one verb (which is further modulated by the syntactic context):

(9) *on se* ***poblil*** *na chodník*
'he threw up on the sidewalk'
(10) *slibte mi, že* ***nepoblijete*** *doktora Reeda*
'promise you will not puke all over Dr Reed'

Of course, if a verb always occurs with *se*, I do not diverge from traditional accounts. One crucial advantage of treating reflexive verbs in this partly unconventional way should become apparent when the causative-inchoative alternation is discussed in section 4.2.

Finally, I distinguished clear lexical ambiguity: e.g., the verb *odsekávat*$_1$ is found in two valency frames (11–12), while the verb *odsekávat*$_2$ is found in one (13):

(11) ***odsekávala***$_1$ ***jsem*** *mu* – 'I kept snapping at him'
(12) *drze* ***odsekáváte***$_1$ *repliky* – lit. '(you) rudely retort lines'
(13) ***odsekávat***$_2$ *maso od kostí* – 'chop the meat from the bones'

I might have distinguished valency patterns that could be lumped together by others, thus, e.g., considering the following examples as instantiations of three different valency frames:

(14) ACT_1^{obl} PAT_4^{obl}
plátky ***napařujte*** *pod pokličkou*
'steam the slices with the lid on'
(15) ACT_1^{obl} PAT_4^{obl} BEN^{obl}
napařuju *si obličej*
'I am steaming my face'

(16) ACT_1^{obl} PAT_4^{obl} $DIR3^{obl}$
na vrstvy polyesterového filmu se ***napařují*** *různé kovy*
'various metals are steamed onto the layers of polyester film'

Other people might conflate (14) and (15) by claiming that both represent the same frame with an optional, albeit typical BEN, which I did not do because while instances similar to (15) are always accompanied by a BEN, instances of (14) in the data never are.

4 SOME OBSERVATIONS

4.1 Czech verbs combine with various valency frames

Clear instances of lexical ambiguity – be it due to polysemy (*vybílit* 'whitewash' → 'clean (out), steal') or homonymy (*nadívat* 'stuff' vs. *nadívat se* 'get enough of looking') – are rather rare in the data, appearing only with 13 out of the 154 verbs. Despite that, each of the 154 plus 13 verbs appears on average with 2.006 valency frames, suggesting that Czech verbs might be more malleable in their valency behavior than presumed. For instance, while *Vallex* gives three valency frames for the relevant senses of the verb *foukat* 'blow,' with the much less frequent verb *profukovat* 'blow through,' we find six valency patterns (some of the corpus examples were shortened):

$DIR2^{typ}$ $DIR3^{typ}$ LOC^{typ}
(17a) *přece jen trochu profukuje* 'it's a bit windy after all'
(17b) *trošku tu profukuje* 'there is a breeze here'
(17c) *okny dovnitř profukuje* 'the wind comes in through the windows'

ACT_1^{obl} $DIR3^{obl}$
(18) *severák profukuje až do kostí* 'the north wind blows through the bones'

ACT_1^{obl} $DIR2^{obl}$ (BEN^{typ})
(19a) *vítr profukuje skulinami* 'the wind blows through the cracks'
(19b) *vítr jí profukoval košilí* 'the wind blew through her shirt'

ACT_1^{obl} PAT_4^{obl} (BEN^{typ})
(20a) *letadlo profukoval ledový vítr* 'an ice cold wind was blowing through the plane'
(20b) *studený vítr profukoval Reedovi bundu* 'a cold wind blew through Reed's jacket'

ACT_1^{obl}
(21) *bunda profukovala* 'the wind was blowing through my jacket' (lit. 'the jacket blew through')

ACT_1^{obl} PAT_4^{obl} (MEANS^{typ})
(22) *profukuje trysku karburátoru* ‘he blows air through the nozzle of a carburetor’

ACT_1^{obl} PAT_4^{obl} DIR2^{obl}
(23) *roztaveným železem se profukoval vzduch* ‘air was blown through molten iron’

While the Functional Generative Description acknowledges, e.g., the systematic alternation between uses such as (19) and (20), it still analyzes these as separate senses of the verb. However, we can claim that all of the uses above are in fact instantiations of a single sense of the verb, and the fact that the verb is interpreted differently in different valency frames does not need to be ascribed to different senses of the verb but to the valency frames, by which we avoid falling prey to the polysemy fallacy mentioned above. Thus, if there is any systematic difference between, say, sentences such as (19) and (20), it can be ascribed to the valency pattern (or, ASC) while maintaining that the meaning of the verb remains the same. Along these lines, the seemingly different meaning of (22) (and of examples such as *okna profukovala* – lit. ‘the windows blew through’) can again be ascribed to the construction rather than the verb itself.

As other examples show, variability in valency behavior is indeed linked to valency alternations that have been described by works such as [23]. However, this is often not the case: various verbs are used in various ASCs without any apparent change in the meaning of the verb even in cases which would not be described as alternations or diatheses, e.g.:

(24) *hlasitě* ***krkne*** – ‘he burps loudly’
(25) ***krknula*** *mi do tváře dvě slova* – ‘she burped two words in my face’

Here one could, indeed, posit one valency frame for the verb (featuring an optional PAT and perhaps an optional ADDR), but this would blur the fact that the (di)transitive use of the verb is marginal (which, however, should never be taken as a reason to discard it), and it does not seem possible to express either the PAT or the ADDR with most of the uses of the verb. The best analysis is in my view the constructionist one, which acknowledges that the verb has only one meaning and in the two examples, it is simply used in two different ASCs. One can posit tentatively the existence of a construction in which an ACT, an ADDR (often not expressed explicitly), and a PAT are required, which coerces the verb into the interpretation of a verbum dicendi, as witnessed by examples (12), (25), and many others, including both common expressions (*řekl mi to* ‘he told me that’) as well as creative uses of verbs, such as *zahalasí nějaký pozdrav* ‘he shouts a greeting,’ in which the verb, usually not used with a direct object, takes one.

4.2 Alternations previously unobserved in Czech

Several of the verbs display what seem to be semantic alternations previously unrecognized in Czech (cf. [23]). In the sample of 154 verbs, two of them, discussed in the following sections, recur across at least a dozen verbs, which suggests that these might be relatively common in Czech. Notably both of them feature especially (albeit perhaps not exclusively) change-of-state verbs. There were other interesting examples of what seem to be previously unrecognized alternations, but these occurred only with one or two verbs, and so for reasons of space, they will not be discussed here.

4.2.1 Causative-inchoative alternation

Multiple verbs in the data allow what is known in the literature as the causative-inchoative alternation [24, pp. 27–30]:

(26a) *pampy se **zazelenaly***
'the pampas turned green'
(26b) *vykrojil z vody souš a **zazelenal** ji bylinami*
'he cut out a patch of land out of the water and turned it green with herbs'
(27a) *...by měla hmota **zesvětlovat***
'the matter should get lighter'
(27b) *Slunce je **zesvětluje**...*
'the sun makes them lighter'
(28a) *směs by měla **napěnit***
'the mixture should foam'
(28b) *my jsme ještě **nenapěnili** mýdlo*
lit. 'we haven't foamed the soap yet'
(29a) *prostě se **přežrala** k smrti*
'she has just eaten herself to death'
(29b) *buď chci nakrmit armádu králíků, nebo se těch pár pokouším **přežrat** k smrti*
'either I want to feed an army of rabbits, or I am trying to make the few eat themselves to death [lit. trying to eat the few to death]'

While in sentences (a) the verb, often (but not necessarily), accompanied by the reflexive morpheme *se*, has an inchoative meaning, in (b) the meaning of the verb is causative, and the PAT of the causative construction[2] corresponds semantically to the ACT of the inchoative construction.[3] Although traditionally these examples would

[2] This is not to be confused with what has been described as the causative diathesis in the Functional Generative Description, illustrated e.g. by *dala*/*nechala dětem spravit boty* 'she had her children's shoes repaired' [25, p. 157].

[3] Examples of this alternation in Czech, although not treated as examples of an alternation, are however mentioned for instance in [26, pp. 223–225] and [27, p. 15].

be treated as featuring two different lexical units (verbs), I believe that both sentences in each pair in fact feature the same sense of the verb, which is, however, further modulated by the ASC in which it is used [11].

Other verbs allowing this alternation include *mutovat* 'mutate'; *vymanévrovat* 'maneuver from'; *rozcinkat* (*se*) '(start to) tinkle'; *vyháknout* (*se*) 'unhook'; *přisouvat* (*se*) 'move, push closer'; *rozesadit* (*se*) 'seat, take seats'; *přetrhat* (*se*) 'break, tear, sever'; *vyplést* (*se*) 'untangle'; *popíchat* (*se*) 'prick' (note that the translations are inevitably somewhat imprecise, especially with respect to Aktionsart).

4.2.2 Agent-Means alternation

Another salient type of alternation previously unrecognized in Czech[4] is one in which a verb takes either an ACT and a PAT, or an ACT, a PAT (corresponding to the PAT in the first configuration), and a MEANS (corresponding to the ACT in the first configuration); cf. [24, p. 80]:

(30a) *lichořešnice rychle* ***ozelení*** *plot*
'the nasturtium quickly covers [lit. makes green] the fence'
(30b) *plot můžeme* ***ozelenit*** *některou z popínavek*
'we can cover the fence with some vines'
(31a) *krev nepřátel se vsakuje do půdy a* ***zúrodňuje*** *ji*
'the blood of the enemies soaks (into) the soil and fertilizes it'
(31b) *bez býložravců, kteří by svými výkaly* ***zúrodňovali*** *půdu…*
'without herbivores which would with their dung fertilize the soil'
(32a) *sprej* ***zohyzdil*** *kašnu* [syn v8]
'the spray damaged the fountain'
(32b) *sprejem* ***zohyzdil*** *fasády* [syn v8]
'he damaged the facades with spray'

Other verbs allowing this alternation include *posilňovat* (*se*) 'strengthen, snack on'; *popíchat* 'prick'; *zahalasit* 'resound'; *zesvětlovat* 'lighten'; *vykurýrovat* 'cure'; *znejisťovat* 'make insecure'; *vystínovat* 'shade'; *ovonět* 'perfume'; *napěnit* 'foam'; *nastřihnout* 'incise'; *rozčleňovat* 'subdivide'; *odbouchnout* 'blow up, shoot' (again, the translations are inevitably somewhat imprecise, especially with respect to Aktionsart).

5 CONCLUSION

Scrutinizing a relatively small sample of verbs of relatively low frequency has shown that the repertoire of valency alternations available in Czech might be richer

[4] Note that in the Functional Generative Description, this alternation is not recognized because of the principle of shifting (i.e., the first participant is always an ACT, irrespective of its semantics).

than previously thought, and that verbs might be somewhat more malleable than is acknowledged by Czech valency theory in that they often seem to combine with multiple valency frames (or, argument structure constructions) without necessarily changing their meaning, much in the spirit of what Construction Grammar has demonstrated for English.

ACKNOWLEDGEMENTS

The preparation of this article was financed within the statutory activity of the Czech Language Institute of the Czech Academy of Sciences (RVO No. 68378092).

References

[1] syn v8: Křen, M. et al. (2019). Korpus SYN [version 8]. Prague: Institute of the Czech National Corpus. Accessible at: www.korpus.cz.
[2] Haugen, T. A. (2013). Adjectival valency as valency constructions: Evidence from Norwegian. Constructions and Frames, 5(1), pages 35–68.
[3] Goldberg, A. E. (1995). Constructions. A Construction Grammar Approach to Argument Structure. Chicago: The University of Chicago Press.
[4] Carlson, G. N., and Tanenhaus, M. K. (1988). Thematic Roles and Language Comprehension. In W. Wilkins (ed.), Thematic Relations, pages 263–288. New York: Academic Press.
[5] de Almeida, R. G., and Manouilidou, C. (2015). The Study of Verbs in Cognitive Science. In R. G. de Almeida and C. Manouilidou (eds.), Cognitive Science Perspectives on Verb Representation and Processing, pages 3–39. Cham: Springer.
[6] Daneš, F., and Hlavsa, Z. et al. (1987). Větné vzorce v češtině. Prague: Academia.
[7] Sgall, P., Hajičová. E., and Panevová, J. (1986). The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Dordrecht: Reidel.
[8] Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. Trends in Cognitive Sciences, 7(5), pages 219–224.
[9] Perek, F. (2015). Argument Structure in Usage-Based Construction Grammar. John Benjamins Publishing Company.
[10] Kemmerer, D. (2015). Cognitive Neuroscience of Language. New York: Psychology Press.
[11] Goldberg, A. E. (2019). Explain Me This. Creativity, Competition, and the Partial Productivity of Constructions. Princeton University Press.
[12] Åfarli, T. A. (2007). Do verbs have argument structure? In E. Reuland, T. Bhattacharya and G. Spathas (eds.), Argument Structure, pages 1–16. Amsterdam: John Benjamins.
[13] Schøsler, L. (2007). The status of valency patterns. In T. Herbst and K. Götz-Votteler (eds.), Valency: Theoretical, Descriptive and Cognitive Issues, pages 51–66. Mouton de Gruyter.
[14] Ickler, I. (2007). Sentence patterns and perspective in English and German. In T. Herbst and K. Götz-Votteler (eds.), Valency: Theoretical, Descriptive and Cognitive Issues, pages 253–269. Mouton de Gruyter.
[15] Borer, H. (2013). Structuring Sense Volume III: Taking Form. Oxford: Oxford University Press.

[16] Lopatková, M. et al. (2020). VALLEX 4.0. Prague: Faculty of Mathematics and Physics, Charles University. Accessible at: https://ufal.mff.cuni.cz/vallex/4.0.
[17] syn2020: Křen, M. et al. (2020). SYN2020: reprezentativní korpus psané češtiny. Prague: Institute of the Czech National Corpus. Accessible at: https://www.korpus.cz.
[18] Zipf, G. K. (1949): Human Behavior and the Principle of Least Effort. Cambridge, Mass.: Addison-Wesley Press.
[19] Divjak, D. (2019). Frequency in Language: Memory, Attention and Learning. Cambridge: Cambridge University Press.
[20] Yi, E., Koenig, J., and Roland, D. (2019). Semantic similarity to high-frequency verbs affects syntactic frame selection. Cognitive Linguistics, 30(3), pages 601–628.
[21] Kettnerová, V., and Lopatková, M. (2010). The Representation of Diatheses in the Valency Lexicon of Czech Verbs. In H. Loftsson, E. E. Rögnvaldsson and S. Helgadóttir (eds.), Proceedings of the 7th International Conference on Advances in Natural Language Processing, IceTAL 2010, pages 185–196. Heidelberg: Springer.
[22] Gillon, B. S. (2015). Optional Complements of English Verbs and Adjectives. In R. G. de Almeida and C. Manouilidou (eds.), Cognitive Science Perspectives on Verb Representation and Processing, pages 67–75. Cham: Springer.
[23] Kettnerová, V. (2014). Lexikálně-sémantické konverze ve valenčním slovníku. Prague: Karolinum.
[24] Levin, B. (1993). English Verb Classes and Alternations. A Preliminary Investigation. The University of Chicago Press.
[25] Urešová, Z. (2011). Valence sloves v Pražském závislostním korpusu. Prague: Institute of Formal and Applied Linguistics.
[26] Kettnerová, V. (2020). Derived lexical reciprocal verbs in Czech. Prace filologiczne, 75(1), pages 215–240.
[27] Veselý, L. (2020). Český vid v kostce. In L. Veselý et al. (eds.), Kapitoly o slovesném vidu nejen v češtině, pages 11–104. Prague: Czech Language Institute of the Czech Academy of Sciences.

sciendo
DOI 10.2478/jazcas-2021-0041

BETWEEN ADVERBS AND PARTICLES: A CORPUS STUDY OF SELECTED INTENSIFIERS

JANA ŠINDLEROVÁ[1] – BARBORA ŠTĚPÁNKOVÁ[2]
[1] Institute of Theoretical and Computational Linguistics, Faculty of Arts, Charles University, Prague, Czech Republic
[2] Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

ŠINDLEROVÁ, Jana – ŠTĚPÁNKOVÁ, Barbora: Between adverbs and particles: A corpus study of selected intensifiers. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 444 – 453.

Abstract: In this paper, we present a preliminary study of three intensifiers (*absolutně*, *naprosto*, *úplně)* based on data from three different corpora, a written corpus SYN2020, a web corpus ONLINE-ARCHIVE, and a spoken corpus ORTOFON 1. Providing a parallel annotation of a random sample of each intensifier, we focus on their functions and meanings in context. We analyse their properties in order to define those features which are relevant to their word class assignment, and to prepare grounds for the future disambiguation tasks.
Keywords: particles, adverbs, intensifiers, corpus, Czech

1 MOTIVATION

"Intensifiers"[1], i.e., words strengthening the meaning of the words in their scope, appear problematic with respect to the word class affiliation of the individual candidate words, i.e., their affiliation either to the class of particles, or to the class of adverbs. So far, the individual studies (including relevant chapters in grammars) concerned with words like *trochu*, *velice*, *úplně* etc. evaluate them differently. This is particularly interesting, because adverbs and particles (as they are defined) should differ both in their syntactic function and in their semantic interpretation. While in Czech linguistic tradition adverbs always work as syntactic constituents, particles do not, they are believed to operate in higher linguistic layers and to acquire pragmatic functions. While adverbs are considered semantic words, having a full meaning, particles are described as synsemantic ones, having a weakened or modal meaning.[2]

[1] We use the term intensifier in this study to avoid referring to all the investigated words in terms of word class categories.

[2] The term *particles* refers here to the category of words expressing the pragmatic dimension of the utterance, as it is traditionally defined in Central European linguistics. The term thus does not apply to words in the function of grammatical operators, e.g., the reflexive element *se* in *dívat se* or *to* in *to be*.

The task of identifying particle uses of intensifier words, and distinguishing them from adverbial uses is needed e.g., for a consistent morphological tagging of linguistic corpora. For example, the disambiguation processes in the current SYN2020 corpus [1] (using the MorfFlex dictionary [2]) almost do not involve particle/adverb rules.

Since the available theoretical studies do not offer a satisfying and thorough argumentation on how to treat intensifying words, a corpus study is needed to describe behaviour of such words and offer corpus-driven criteria to support their word-class categorization.

In this paper, we present a preliminary corpus study aimed at three representative intensifying words: *absolutně, naprosto,* and *úplně* ('absolutely, completely, totally'). In their lexicographic treatment, they are often presented as near synonyms, expressing similar meanings and appearing in similar contexts, cf. SSČ [3]. We investigate their function, their meaning, and their context in three different corpora, and based on a pilot annotation, we point out features leading to difficulties in the task of word class disambiguation.

2 PARTICLES VERSUS ADVERBS: THE THEORY

There is a considerable lack of criteria to delimit the category of particles as a unified and compact system. The existing criteria are largely negative in nature. Particles are primarily a) inflective, b) synsemantic, and c) they do not function as a clause constituent (see e.g., [4, p. 90]). A further delimitation of particles in contrast to other synsemantic word classes works on the basis of elimination: they do not assign case, they do not conjoin words or clauses. As category-unifying features, mostly the following are presented: ability to modify clauses, ability to link the proposition with the context, and expressing the relation of the speaker to the communication situation (cf. e.g., MČ2 [5, p. 228] or VAGSČ [6, p. 91]). There are attempts to define semantically compact subclasses within the category, nevertheless, the individual authors differ in the number and extent of the subclasses distinguished.

One of the subclasses sometimes identified within the class of particles are the intensifiers, i.e., words like *velmi, zcela, úplně*. Before establishment of particles as a separate category, they were generally considered adverbs. Moreover, some of them also still hold a separate adverbial meaning (e.g., they can be used as obligatory adverbial complementations of verbs, etc.). Therefore, they are sometimes treated as adverbs of measure in literature.

MČ2 [5] treats intensifiers under the label of measure (or intensification) adverbs, which is considered a subclass of manner adverbs. The measure adverbs are given as a list, without detailed characterization or contextual exemplification (p. 190). Additionally, intensifiers are treated in the chapter of particles as well, this time as a subclass of "measure evaluating particles". Again, the potential members

of the class are listed only and the list overlaps with measure adverbs. A similar approach is offered in PMČ [7].

It is quite clear that even when looking into a single grammar of the Czech language, it may not be obvious where the borderline between two distinct uses of the same intensifier is. The underlying cause of the ambiguous treatment of the so-called measure adverbs and intensifying particles is the problem of defining a difference between intensification and emphasis, a subject broached in the Czech linguistic discourse as early as starting with Mathesius (1947; [8]), who acknowledges their overlapping character. Nekula (PMČ [7, p. 360]) mentions that in some cases, the word class affiliation can be influenced by the relative position of the intensifier to the affected word. Vondráček in VAGSČ [6, p. 103] summarizes his findings from a separate study on the topic [9] in the following way: the criterion for distinction of the two word classes lies in the dominance of either the function of specifying measure, or the generally modifying function, together with the ability of the intensifier to work as a syntactic complementation. A similarly general solution is offered by Šimková (2002; [10]).

Looking at the lexicographic approaches to word class assignment to intensifiers, the older monolingual dictionaries SSČ [3] and SSJČ [11] usually assign a single label, either a particle, or an adverb. An attempt to differentiate between the two word classes through separate entries can be seen in the newly prepared ASSČ [12] on the example of *absolutně*, which is presented as an adverb, and also a particle.

3 DATA ANNOTATION

3.1 Annotation process

As a data resource, we have chosen three different corpora representing three different types of text: a representative written corpus SYN2020 [1], a web corpus ONLINE_ARCHIVE [13] and a representative corpus of spoken language ORTOFON [14]. For each of the selected intensifiers (*absolutně, naprosto, úplně*), we have obtained 50 random concordances.[3]

Each of the intensifiers was annotated for the following features: the word class of the word in the scope of the intensifier; the position of the intensifier relative to the word in its scope; the function of the affected word within the clause; the position of the intensifier within the clause; the intensifier's assumed word class.

Most concordances included an intensifier modifying an adjective or a verb, consequently, the most frequent syntactic functions of the word in the scope were verbonominal predicates, verbal predicates, and attributes. Majority of cases were intensifiers in the anteposition to the affected words, most prominently in

[3] The intensifier *absolutně* gives only 45 concordances in the ORTOFON corpus overall.

a penultimate position in the clause. The spoken corpus showed a notable number of intensifier postpositions, related to a higher number of clause-final positions.

The annotation of syntactic features also served the purpose of making sure that both annotators interpreted the meaning of the sentence in the same way, which is especially important in the case of web and spoken corpora. The results showed that clearly different interpretations were indeed rare.

3.2 Inter-annotator agreement

We calculated the inter-annotator agreement simply as a proportion of the number of cases where the annotators agreed on the label assigned to the number of cases where they differed in the assigned labels. For this pilot study, we were interested mainly in the overall certainty of the annotators to assign the labels and whether the points of disagreement share some common semantic or syntactic features.

Points of disagreement in the word class affiliation of the affected words were infrequent and concerned mainly delimitation of the scope in the case of verbonominal predicates or complex predicates. This often resulted in disagreement in the syntactic function of the affected complementation. None of the differences in judgement of morphosyntactic properties of the affected word seems to have had a direct impact on the word class assignment to the intensifier itself.

Our main interest lied in the agreement on the word class categorization of the intensifier itself. We hypothesized that in view of the fact that the current linguistic theories do not offer a satisfying account of what constitutes a particle, the inter-annotator agreement in this task would be rather low. This was confirmed by the data. In the SYN2020 [1] and ONLINE [13] data, the agreement was lower than 75 %; the annotators chose different labels in more than a quarter of the occurrences.

Intensifier	*absolutně*			*úplně*			*naprosto*		
	D	**T**	**DA**	**D**	**T**	**DA**	**D**	**T**	**DA**
SYN2020	10	32	6	19	11	20	31	3	16
ONLINE	7	34	9	22	13	15	25	6	17
ORTOFON	8	31	6	10	28	10	23	19	8

Tab. 1. Annotator agreement[4]

Only a few, syntactically restricted types of contexts allow a definite agreement on the word class assignment. E.g., if the intensifier modifies a syntactic noun

[4] We use D for adverbs, T for particles, and DA for disagreement. The numbers in the tables in some cases do not add up to the total number of concordances evaluated. We did not include into the overall counts cases when the word class assignment was not possible due to the utterance being too fragmented.

(nouns, totalizing or indefinite pronouns, rarely an infinitive) in the function of a syntactic subject or object, the intensifier is annotated as T.

The word class affiliation of adjectives (whether in the syntactic function of an attribute, or as the nominal part of the verbonominal predicate) is rather unconditioned by the morphological and syntactic interpretation of the word in context. The decision is thus probably driven by semantic factors. This is rather bad news for the efforts to formulate contextual disambiguation rules for automatic processes.

The fact that the same intensifiers might be assigned both the adverb label and the particle label in similar contexts, similar clause positions, and similar syntactic constructions indicates that the process of word class recategorization (from adverb to particle) has not been yet completed. Nevertheless, the annotation and the following analysis show certain semantic tendencies.

A significantly higher agreement in the *absolutně* data suggests that the process of particulization of the expression is already advanced.

The agreement is considerably better in the spoken corpus data. Also, in spoken data, the particle label (T) is more frequently assigned considering the otherwise ambivalent intensifiers *naprosto* and *úplně*. This is probably caused by the fact that spoken data tend to include more clear attitude markers than the written ones.

3.3 Selected points of disagreement in the intensifier word class

The certainty of the annotators varied regarding each individual intensifier. Whereas *absolutně* achieves the highest agreement (85 %), *naprosto* and *úplně* reach around 70 %. While the highest agreement was observed in the spoken corpus, it seems that the disambiguation issue may be supported by the prosodic properties of the utterance, or by other features of a spoken text (simple structures etc.).

Typically, the disagreement arises in combination with evaluative expressions, with qualitative adjectives and with positive verb forms.

By evaluative expressions, we mean primarily qualitative qualification adjectives or adverbs expressing subjective evaluation (1), cf. Karlík in NESČ [15].

(1) *Školním divadlem jsem byla naprosto nadšená.* 'I was totally excited about the school theatre performance.'

Here, the annotation basically confirms Vodráček's opinion [9] that distinguishing adverbial from particle meaning is often dependent on "the semantic interpretation of the modified or specified expression", i.e., whether the intensifier in context expresses measure circumstance or speaker's attitude. Nevertheless, our annotation data suggest that the decision is rather subjective and based probably on whether the annotator perceives the expression in the given context as scalable or as a representation of an upper limit which is only emphasized, see an example of a disagreement in (2).

(2) *to je úplně jednoduchý* 'That's totally simple.' (ORTOFON)

Whereas *absolutně* modifying a negative verb is assigned a T label almost uniformly, positive verb forms trigger both measure interpretation (intensification) and emphasizing interpretation (3).

(3) *absolutně mě dokáže odradit chlap, který je blbý* 'I feel absolutely appalled by a guy who is dull.' (ONLINE)

Naprosto and *úplně* seldom affect negated verbs, therefore, the disagreement concerns mainly combinations with positive verb forms.

4 INTENSIFIERS IN CONTEXT

4.1 Absolutně

The word *absolutně* appears to be strongly tied to negative contexts. By negative contexts, we mean collocations with explicit morphological negation (4), as well as collocations with words and phrases with negative meaning but no morphological marking (5).

(4) *...absolutně jsem netušil, jak se bude můj život dál vyvíjet*. 'I had absolutely no idea how my life would go on.' (SYN2020)

(5) [Je] *Absolutně vyloučeno, abych zabloudil.* '[It´s] Absolutely out of question for me to get lost.' (SYN2020)

The tendency to be bound to negative context is the strongest in ORTOFON (40 out of 45), the corpus of informal spoken Czech, (42 out of 50) while in SYN2020, a corpus of written texts, the number of positive collocations rises and negative contexts reach only 31 out of 50. This fact constitutes one of the major function differences between *absolutně* on one hand and *úplně* or *naprosto* on the other hand, making it close to *vůbec*, an accepted Czech "negative polarity item".[5]

The positive contexts include the following types: the intensification of evaluative adjectives or adverbs (6), the intensification of words expressing sameness or different character (*stejný, jinak*) (7), or modification of an objective quality (8).

(6) *Auto, které vidíte na obrázku před sebou se jmenuje Interceptor S a je absolutně skvostné.* 'The car you see in the picture in front of you is called Interceptor S and is absolutely brilliant.' (SYN2020)

[5] *Absolutně*, though, cannot be considered an NPI, since it fits perfectly into positive contexts: *S názvem článku absolutně nesouhlasím.* 'I absolutely do not agree with the article title.' *S názvem článku absolutně souhlasím.* 'I absolutely agree with the article title.'

(7) *Půjde o absolutně jiný film.* 'It will be an absolutely different film.' (SYN2020)

(8) *Stal jsem se na ní absolutně závislej.* 'I became absolutely dependent on her.' (SYN2020)

Rarely, the word absolutně appears in the function of a clear manner adverbial (9).

(9) *Pokud řada nekonverguje absolutně, může její součet být v rozporu s naším očekáváním.* 'If a series does not converge absolutely, its sum may be contrary to our expectation.' (SYN2020)

As for the position of the word within a sentence, while in SYN2020 and ONLINE, *absolutně* stands almost exclusively in front of the word in its scope, sometimes in the middle position (usually with copular predicates), the ORTOFON data show 7 cases of a postposition (10). In case the sentential stress lies on the verb in the scope, the word is then easily interpreted as a particle.

(10) *Jo to já ti rozumím absolutně*. 'Yeah, I absolutely understand you.' (ORTOFON)

4.2 Naprosto

In contrast to *absolutně*, *naprosto* does not appear significantly in negative contexts, though a negative context is not excluded (11).

(11) *Zatím bohužel rozšiřuje řady kočiček, kterým útulek naprosto nevyhovuje*. 'Unfortunately, it is increasing the number of cats, for which a shelter is not convenient at all.' (ONLINE)

Naprosto often modifies positive evaluative words (12), but it also appears with words describing sameness or different character (13).

(12) *Vše se nese v naprosto pohodové atmosféře.* 'Everything is carried on in a totally relaxed atmosphere.' (ONLINE)

(13) *Barevnost nechali čistě na nás, měli jsme však vybrat naprosto odlišné odstíny, než byly v původní ložnici.* 'They left the color purely on us, but we had to choose totally different shades than those of the former bedroom.' (SYN2020)

Also, in comparison to the other two intensifiers discussed, *naprosto* is notably more often used even with non-evaluative words (i.e., words expressing some quality primarily without judging whether the quality is positive or negative) (14).

(14) *Vše tkví ve výchově a vzdělání a v těchto dvou oblastech si mohou být ženy s muži naprosto rovny.* 'It's all about upbringing and education, and women can be totally equal to men in these two areas.' (SYN2020)

4.3 Úplně

Úplně, in contrast to the previous candidates, almost does not appear with negative verbs in our data. This is probably connected to the fact that when combined with a negated verb, its meaning is shifted. While *naprosto* expresses an upper limit, a total completion of the verb meaning with positive verbs (15a), and a lower limit, a total incompletion of the verb meaning with negative verb (15b), *úplně* expresses an upper limit, a total completion of the verb meaning with positive verbs (15a), but with negative verbs in some syntactic contexts, its meaning is rather reaching a low level of completion of the verb meaning (15c).

(15a) *Naprosto/Úplně s tebou souhlasím.* 'I strongly agree with you.'

(15b) *Naprosto s tebou nesouhlasím.* 'I strongly disagree with you.'

(15c) *Úplně s tebou nesouhlasím.* 'I slightly disagree with you.'

Úplně in this context weakens the negative meaning, rather than intensifying it.

In positive contexts, *úplně* more or less shares the usual collocations of *naprosto* and *absolutně*. It combines with the sameness or difference expressions more often than *naprosto* (16).

(16) *tam je úplně jiná mentalita ještě* 'there is a completely different mentality yet'(ORTOFON)

It combines with positive totalizing pronouns (*všechno*, *každý*), but negative pronouns are extremely rare in its scope, mostly they are perceived as rather incompatible, because the natural opposite to *úplně* in negative contexts is *vůbec* (17).

(17) *Neměl jsem vůbec/??úplně žádné peníze*. 'I had no money at all.'

We can find examples of *úplně* modifying evaluative words, though they are less frequent than the occurrences of evaluative words after *naprosto*.

Úplně seems to combine easily with descriptive adjectives expressing a neutral quality (18).

(18) *Je úplně svěží a čeká.* 'She is completely fresh and waiting.' (SYN2020)

Last, but not least, *úplně* appears in the meaning of "sort of", "almost" (19).

(19) *Úplně se mi sbíhají sliny.* 'My mouth is almost watering.' (ONLINE)

5 CONCLUSIONS

We have carried out the data annotation and analysis of selected intensifiers in order to check out whether human annotators are able to provide consistent and reliable disambiguation decisions based on the available definitions of the adverb and particle categories. Also, because a detailed explanation of the different uses of intensifiers, supported by strong evidence from real data, is usually missing in the current grammars, we wanted to identify important features and properties of the intensifier uses in context that would possibly help shed light on the disambiguation process.

The three intensifiers investigated do not behave in the same way in the context. *Absolutně* is extremely likely to appear in negative contexts, whereas *úplně* acquires a specific meaning with negation in its scope. The annotation suggests that the shift from understanding the intensifier as an adverb to considering it a particle is most advanced with *absolutně*, while *naprosto* is still interpreted in the sense of an adverb, and *úplně* maintains both interpretations. This may also explain the fact that the inter-annotator agreement was the lowest with *úplně*.

Only a few, syntactically restricted types of contexts, allow a definite agreement on the word class assignment, such as intensification of nouns and syntactic nouns, above all totalizing or negative pronouns.

The word class affiliation of intensifiers affecting evaluative expressions is rather unconditioned by the morphological and syntactic interpretation of the affected word. In many cases, the decision is probably driven by semantic factors, above all by the subjective interpretation of its semantics as scalable or non-scalable. This impact of subjective evaluation is a serious factor hindering the efforts to formalize the disambiguation task for automatic analysis.

ACKNOWLEDGEMENTS

This paper has been in part supported by the LINDAT/CLARIAH-CZ project funded by Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101).

This paper has been in part supported by the Ministry of Education of the Czech Republic, through the project Czech National Corpus, no. LM2018137.

References

[1] Křen, M. et al. (2020). SYN2020: reprezentativní korpus psané češtiny. ÚČNK FF UK, Praha. Accessible at: http://www.korpus.cz.

[2] Hajič, J. et al. (2020). MorfFlex CZ 2.0. Data/software, LINDAT/CLARIAH-CZ digital library, Czech Republic. Accessible at: http://hdl.handle.net/11234/1-3186.
[3] Filipec, J. et al. (2003). Slovník spisovné češtiny pro školu a veřejnost. Praha. [SSČ].
[4] Čechová, M. (2000). Čeština – Řeč a jazyk. Praha.
[5] Komárek, M. et al. (1986). Mluvnice češtiny 2. Praha. [MČ2].
[6] Štícha, F. et al. (2018). Velká akademická gramatika spisovné češtiny. Praha. [VAGSČ].
[7] Karlík, P. et al. (1995). Příruční mluvnice češtiny. Praha. [PMČ].
[8] Mathesius, V. (1947). Zesílení a zdůraznění jako jevy jazykové. In Čeština a obecný jazykozpyt, pages 203–223, Praha.
[9] Vondráček, M. (1999). Příslovce a částice – hranice slovního druhu. Naše řeč 82, pages 72–78.
[10] Šimková, M. (2002). Príslovky a častice *celkom*, *úplne – vôbec*. In Varia 9, pages 325–328, Bratislava.
[11] Havránek, B. et al. (1960–1971). Slovník spisovného jazyka českého. Praha. [SSJČ].
[12] Akademický slovník současné češtiny. (2017–2020). Praha. [ASSČ]. Accessible at: https://slovnikcestiny.cz.
[13] Cvrček, V., and Procházka, P. (2020). ONLINE_ARCHIVE: monitorovací korpus internetové češtiny. ÚČNK FF UK, Praha. Accessible at: http://www.korpus.cz.
[14] Kopřivová, M. et al. (2017). ORTOFON, verze 1 z 2. 6. 2017. ÚČNK FF UK, Praha. Accessible at: http://www.korpus.cz.
[15] Karlík, P. (2017). Adjektivum. In P. Karlík et al. (eds.), CzechEncy – Nový encyklopedický slovník češtiny. Accessible at: https://www.czechency.org/slovnik/ADJEKTIVUM.

sciendo
DOI 10.2478/jazcas-2021-0042

CAPTURING NUMERALS AND PRONOUNS AT THE MORPHOLOGICAL LAYER IN THE PRAGUE DEPENDENCY TREEBANKS OF CZECH

BARBORA ŠTĚPÁNKOVÁ – MARIE MIKULOVÁ
Charles University, Prague, Czech Republic

ŠTĚPÁNKOVÁ, Barbora – MIKULOVÁ, Marie: Capturing numerals and pronouns at the morphological layer in the Prague Dependency Treebanks of Czech. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 454 – 464.

Abstract: The paper presents a novel and unified morphological description of numerals and pronouns, as compiled for the newest edition of the Prague Dependency Treebank (Prague Dependency Treebank – Consolidated 1.0) and its integral part the morphological dictionary MorfFlex. On the basis of considerable experience with real data annotation and the use of the morphological dictionary, particular changes were proposed. For both of the parts of speech a new set of subtypes was proposed, based mainly on the morphological criterion and its combination with semantic properties and other relevant features, such as definiteness in numerals and possessivity, reflexivity, and clitichood in pronouns. Each subtype has a specific value at the 2nd position of the morphological tag, which serves also as an indicator of the applicability of other tag categories.

Keywords: numerals, pronouns, morphology, treebank, annotation, Czech

1 INTRODUCTION

The paper is focused on numerals and pronouns and their morphological description in the Prague Dependency Treebank. Although these word categories are considered to be a traditional part of the part of speech (POS) set in Czech linguistics, their morphological (as well as some other) properties are quite specific and different from the rest of the inflected words. For the newest edition of the Prague Dependency Treebank (Prague Dependency Treebank – Consolidated 1.0 [1], PDT-C in the sequel) we compiled a novel and unified description of these two POS and proposed its realization in the morphological tag. The PDT-C 1.0 release is enhanced with a manual linguistic annotation at the morphological layer: all tokens of the sentence are tagged and lemmatized. A key element to ensuring annotation consistency is the morphological dictionary MorfFlex [2]. Based on the long-time experience with the use of the dictionary and manual annotation of real data, some phenomena were proposed to be captured differently in the dictionary in order to achieve better consistency within the dictionary as well as between the dictionary and the annotated data. The changes concern several complicated morphological features of Czech, including some relating to numerals and pronouns.

2 NUMERALS AND PRONOUNS IN CZECH GRAMMARS

In grammars, a POS is usually defined on the basis of a combination of semantic, morphological, and syntactic criteria. Numerals and pronouns are mainly delineated by appeal to the semantic criterion, while the other two criteria are rather problematic with these two classes.

Numerals semantically indicate number or quantity (e.g., *pět* 'five', *třetí* 'third', *několik* 'several'). However, in various approaches the set of words included as numerals can differ, e.g., *pětina* 'a fifth' is often considered a noun, despite its numeric meaning and digit representation.[1] The overarching semantic feature of pronouns lies in their ability to substitute for nouns or adjectives (e.g., *on* 'he', *tento* 'this', *nějaký* 'some').

From the morphological point of view, and in many respects also from the syntactic point of view, both numerals and pronouns can also be considered as different POS (e.g., *tisíc* 'thousand' – numeral/noun; *který* 'which' – pronoun/adjective; *druhý* 'second' – numeral/adjective; *mnohokrát* 'many times' – numeral/adverbial). On the other hand, words like *někde* 'somewhere' and *kdykoliv* 'anytime' are usually regarded as pronominal adverbs.[2] For these reasons, the most recent Czech grammar [5] does not distinguish pronouns and numerals as individual POS and assigns them among other POS.

3 NUMERALS AND PRONOUNS IN PDT-C

Within the morphological annotation in PDT-C, forms are organized into entries (paradigms) according to their formal morphological behavior. The paradigm (set of forms) is identified by a unique lemma. For each form, full inflectional information is coded in 15 tag positions. The first two tag positions encode the part of speech. The traditional POS is captured in the 1st position. The 2nd position of the tag specifies the detailed subtype of the POS and serves as an indicator of the (non-)applicability of the other categories encoded in the tag [7, p. 17]. The other positions of the tag capture morphological properties, of which gender (3rd position), number (4th), case (5th), possgender (6th), possnumber (7th) and variant (15th) (explained in 5.3 and 5.4) are relevant to the description of pronouns and numerals (cf. [6] and [7]).

The PDT approach to numerals and pronouns follows the traditional classification, i.e., numerals and pronouns are considered separate POS (the 1st tag position has the value P in the case of pronouns and C in the case of numerals). In the next part of the article, we concentrate on the description of the 2nd tag position. The main criteria for a more detailed classification of numerals and pronouns are semantics and morphological behavior. We include the semantics as a criterion for traditional reasons. In linguistics, numerals and pronouns are traditionally classified according to their

[1] In the NovaMorph morphological description [3] words of this type are included in a group of fractional numerals.

[2] Komárek united these semantically similar words into one group of "deictic words" [4].

purpose (as e.g., cardinal, ordinal, multiplicative, in the case of the former, and as e.g., personal, demonstrative, relative, in the case of latter).

However, the crucial principle of the classification at the 2nd tag position is the morphological one. As already mentioned, the value of the 2nd position serves as an indicator of the applicability of the other tagged categories (such as gender, number, and case) and this principle is decisive for the system of pronouns and numerals in PDT-C.[3] This means that, for example, a numeral expressing agreement in grammatical gender (e.g., *jeden, jedna, jedno* 'one') cannot be in the same group as a numeral that behaves as a noun in that it has a single, fixed gender (e.g., *sto* 'hundred') or does not express any gender (e.g., *tři* 'three') although according to the semantic criterion, such numerals may belong to the same category (the numerals *jeden*, *sto* and *tři* are all cardinal numerals; cf. Tab. 1).

In each POS group, the two main criteria (semantic and morphological) are supplemented by other criteria, following from the properties of the group. Within the class of numerals, there is the criterion of definiteness, and within the group of pronouns, possessivity, reflexivity, and clitichood are considered.[4]

4 SUBTYPES OF NUMERALS

Numerals (except for the numbers written down with Arabic or Roman numerals) are classified into several subtypes according to the various combinations of the two basic features (cf. Tab. 1):

- morphological behavior,
- semantics including definiteness.

Semantics & Morphological behavior		**Adjectival**		**Nominal**		**Non-gendered**	
Type	Subtype	Tag	Example	Tag	Example	Tag	Example
Cardinal	definite	**Cn**	*jeden*	**Cz**	*sto*	**Cl**	*tři*
	indefinite	**Cy**	*nejeden*	-	-	**Ca**	*málo*

[3] In the previous proposal [6], this principle was violated in several cases. The new proposal eliminates these violations.

[4] By comparison, some other West-Slavonic languages corpora approach description of numerals and pronouns differently. The description in the Slovak National Corpus [8] is based on a morphological criterion, the primary classification corresponds to the declension type (nominal, adjectival, mixed, etc.) that indicates other tag positions. The semantic criterion is present by inclusion of deictic words among pronouns (cf. [9]). The National Corpus of Polish takes into account mainly the morphological and syntactic criterion (strong vs. non-strong position, post-prepositional vs. non-post-prepositional position). The semantic criterion is noticeable only in a few tag positions (e.g., different values for main vs. collective numerals, personal vs. personal reflexive pronouns; cf. [10] and [11]).

Semantics & Morphological behavior		Adjectival		Nominal		Non-gendered	
Ordinal	definite	**Cr**	*první*	-	-	**-**	-
	indefinite	**Cw**	*kolikátý*	-	-	**-**	-
Multiplicative	definite	**-**	-	-	-	**Cv**	*třikrát*
	indefinite	**-**	-	-	-	**Co**	*tolikrát*
Collective	definite	**Cd**	*dvoje*	**Cj**	*patero*	**-**	**-**
	indefinite	**Ch**	*kolikerý*	**Ck**	*kolikero*	**-**	**-**
Arabic	-					**C=**	*1,25*
Roman	-					**C}**	*MDX*

Tab. 1. Subtypes of numerals

4.1 Morphological behavior

The morphological criterion is fundamental. We distinguish between adjectival, nominal and non-gendered numerals (cf. Tab. 1).

Form	Lemma	Tag
jeden muž	jeden`1	`CnYS1----------`
jedno dítě	jeden`1	`CnNS1----------`
několikátý problém	několikátý	`CwYS1----------`
několikáté problémy	několikátý	`CwIP1----------`

Tab. 2. Examples of adjectival numerals

The **adjectival numerals** express the same gender (and also the number) as that of the governing noun (e.g., *jedna žena* 'one woman' (fem. sg.), *jedno dítě* 'one child' (neut. sg.) or *několikátý problém* 'umpteenth problem' (masc. sg.), *několikáté problémy* 'multiple problems' (masc. pl.)). All forms are represented by one lemma; similarly to adjectives. However, in contrast to adjectives there is no comparative and superlative form in the adjectival declension of numerals, and some paradigms have only singular (e.g., *jeden* 'one'), or only plural forms (e.g., *dva* 'two') depending on their meaning. In the tag, the gender, number, and case are specified (see Tab. 2).

The subtype of numerals with **nominal declension** consists of numerals, whose morphological behavior is similar to that of nouns (e.g., *sto* 'hundred', *nula* 'zero', *patero* 'five-kinds-of'). They express grammatical gender, thus their tag position for gender is filled, with the whole paradigm sharing the same value (see Tab. 3).

Form	Lemma	Tag
sto lidí	sto-1`100	CzNS1----------
napsal dvojku a tři *nuly*	nula	CzFP4----------
z *patera* přikázání	patero`5	CjNS2----------

Tab. 3. Examples of nominal numerals

Other numerals express **no gender** and they are quite specific and diverse from the morphological point of view; they can be inflected (e.g., *tři domy* (Nom.), *tří domů* (Gen.) 'three houses'; *mnoho domů* (Nom.), *mnoha domů* (Gen.) 'many houses'), or they are uninflected, sometimes with a variant form (e.g., *několikrát, několikráte* 'several-times').[5] The tag position for gender is not filled. Other tag positions (case, number) are filled if the subtype expresses them (cf. Tab. 4).

Only the group of definite cardinal numerals covers all morphological subtypes, i.e., it expresses agreement in grammatical gender (then the tag begins with Cn; e.g., *jeden* 'one'), only one (lexical) gender (then the tag begins with Cz; e.g., *sto* 'hundred') or it expresses no gender (then the tag is Cl; e.g., *tři* 'three'); cf. Tab. 1.

Form	Lemma	Tag
tři domy	tři`3	Cl-P1----------
do *mnoha* zemí	mnoho	Ca--2----------
několikrát zazvonil	několikrát	Co----------
několikráte zazvonil	několikrát	Co----------1

Tab. 4. Examples of non-gendered numerals

In certain contexts, numerals do not mark grammatical relations with other words in a sentence by means of inflection, they are used in their uninflected form. In the following examples, despite the formal differences (and some stylistic nuances) the meaning of these variant forms is broadly equal, i.e., the inflected forms of *sto* 'hundred' in *ke čtyřem stům dětem* 'up to four hundred children' or *do sta lidí* 'up to a hundred people' can also be expressed with the uninflected form, as in *ke čtyři sta dětem* 'up to four hundred children' and *do sto lidí* 'up to a hundred people'. In particular combinations, uninflected forms are quite usual, e.g., *až po stovky tisíc let* 'up to hundreds of thousands of years'; *sedmdesát jedna občanů* 'seventy-one citizens'; *před tři čtvrtě rokem* 'three quarters of a year ago'; *o pár stech tisících* 'about a few hundreds of thousands'. Thus, for most cardinal numerals we introduced the subspecified value X for any gender, number, or case; cf. Tab. 5.

[5] Variant forms are distinguished at 15th tag position by a numerical index (cf. Tab. 4).

Form	Lemma	Tag
ke *čtyřem* stům dětem	čtyři`4	`Cl-P3----------`
ke *čtyři* sta dětem	čtyři`4	`Cl-XX----------`
ke čtyři *sta* dětem	sto-1`100	`CzNXX---------1`
do *sto* lidí	sto-1`100	`CzNXX----------`
sedmdesát *jedna* občanů	jeden`1	`CnXXX----------`
před tři *čtvrtě* rokem	čtvrt	`CzFXX---------1`

Tab. 5. Examples of uninflected forms of numerals

4.2 Semantics including definiteness

In accordance with semantic features, we classify numerals into the following subtypes. Within the each semantic subtype, the numerals are further classified according to their definiteness; cf. Tab. 1.

Cardinal – express quantity (e.g., *oba* 'both', *kolik* 'how much', *málo* 'a little', *milion* 'million').

Ordinal – express position in a sequential order (e.g., *třetí* 'third', *několikátý* 'umpteenth').

Multiplicative – express how many times something occurred (e.g., *sedmkrát* 'seven-times', *mnohokrát* 'many-times').

Collective – express the number of kinds, types (e.g., *dvojí* 'two-kinds-of', *několikerý* 'several-kinds-of', *desatero* 'ten-kinds-of').

In contrast to Czech grammars, we do not distinguish interrogative numerals as a separate type: interrogative numerals are included in the corresponding types of indefinite numerals; e.g., *kolik* 'how-many' is included in the cardinal indefinite type, *kolikátý* 'at-what-position-in-a-sequence' is included in the ordinal indefinite type, or *kolikrát* 'how-many-times' is included in the multiplicative indefinite type.

5 SUBTYPES OF PRONOUNS

Pronouns form a more complicated part of speech than numerals. More criteria need to be considered for their adequate classification. As with the category of numerals, we chose the morphological behavior and semantics as the main criteria for pronouns.

Then we identified several other features that can be used to divide pronouns into various subtypes (cf. Tab. 6):

- possession,
- reflexivity,
- clitichood.

5.1 Morphological behavior

The morphological criterion divides pronouns into two groups: gendered pronouns and non-gendered pronouns.

Semantics & morphological behavior		**Gendered**		**Non-gendered**	
Type	Subtype	Tag	Example	Tag	Example
Personal	- Clitic	**PE** **P5**	*on, něj* *mu*	**PP** **PH**	*já, ty, vy* *mi*
	Reflexive Reflexive Clitic	- -	- -	**P6** **P7**	*sebe* *se, si*
	Possessive	**PS**	*můj, náš*	-	-
	Possess. 3rd Pers	**P9**	*jeho, jejich*	-	-
	Reflex. Possess.	**P8**	*svůj*	-	-
Relative	-	**P4**	*který, čí, jenž*	**PQ**	*kdo, copak*
	Possessive	**P1**	*jehož, jejichž*		
Indefinite	-	**PZ**	*nějaký, čísi*	**PK**	*kdosi, nevímco*
Negative	-	**PW**	*nijaký, žádný*	**PY**	*nikdo, nic*
Demonstrative	-	**PD**	*ten, takový*	-	-
Delimitative	-	**PL**	*všechen*	-	-

Tab. 6. Subtypes of pronouns

Gendered pronouns express different values of the gender (and also number):

(a) depending on the grammatical gender and number of the governing noun; cf. *žádný dům* 'no house' (masc. inanim. sg.), *žádná žena* 'no woman' (fem. sg.), *žádní muži* 'no men' (masc. anim. pl.). These pronouns behave as syntactic adjectives in sentences;

(b) according to the gender, animacy or number of the referent they substitute (e.g., *on* 'he', *ona* 'she', *ono* 'it', *oni* 'they'), they behave syntactically as nouns.

All forms of both of the types are represented by one lemma (Nom. sg. masc. anim.), similarly to adjectives: e.g., *žádný* 'no' for *žádná* (fem.), *žádné* (neut.), and *on* 'he' for *ona* 'she' and *oni* 'they', etc. The tag positions for gender and number are filled; cf. Tab. 7.

Non-gendered pronouns are pronouns that express no gender and number (e.g., *ty* 'you'; *nikdo* 'nobody', *cosi* 'something'; Tab. 8). The gender and number tag positions are not filled. These pronouns behave syntactically as nouns.

Form	Lemma	Tag
on	on-1	PEYS1--3-------
oni	on-1	PEMP1--3-------
ono	on-1	PENS1--3-------
žádná žena	žádný	PWFS1----------
žádní muži	žádný	PWMP1----------

Tab. 7. Examples of gendered pronouns

Form	Lemma	Tag
nikdo	nikdo	PY--1----------
nikoho	nikdo	PY--2----------
cosi	cosi	PK--1----------
čehosi	cosi	PK--2----------

Tab. 8. Examples of non-gender pronouns

We are aware that the pronoun *kdo* 'who' (and other various personal pronouns) could be classified as masculine, and the pronoun *co* 'what' (and other various non-personal pronouns) could be classified as neuter. However, there are many uses of these pronouns where the gender and number category seems to be questionable (e.g., *kdo jste tam byli* (masc. anim. pl.) 'who of you were there', *každá* (fem. sg.)*, kdo jste přišla* (fem. sg.) 'each of you who came', *nikdo nejsme* (1st pers. pl. – 'we') *dokonalý* 'none of us are perfect'; cf. also examples and the discussion in [12]).

5.2 Semantics

We identified six main semantic groups of pronouns, largely following the Czech grammar tradition. In each group, we use the unique tag value to distinguish between gendered and non-gendered pronouns:

Personal: substitute a particular word referring to a person, thing, and the like, including pronouns indicating clitichood, possession and reflexivity (see below); e.g., *já* 'I', *on* 'he', *ní* 'her', *náš* 'our', *jeho* 'his', *svůj* 'self'.

Relative/Interrogative: in relative clauses, they are used as connecting words referring back to their antecedents; in questions, they serve as interrogative words; e.g., *jaký* 'what', *který* 'which', *čí* 'whose', *co* 'what', *kdož* 'who'.

Indefinite: refer to one or more unspecified persons or things; e.g., *nějaký* 'some', *čísi* 'somebody's', *sotvakterý* 'hardly-some', *někdo* 'somebody', *kdokoliv* 'whoever'.

Negative: refer to nonexistence of persons, things or their properties; e.g., *ničí* 'nobody's', *žádný* 'no/none', *nic* 'nothing', *nikdo* 'nobody'.

Demonstrative: point to a specific person or thing; e.g., *ten* 'this', *tamten* 'that', *onen* 'that-over-there', *tentýž* 'same', *takový* 'such'.

Delimiting: (sometimes included in the indefinite group [13, p. 224]) indicate the universality or totality; e.g., *všechnen* 'all', *sám* 'alone', *veškerý* 'whole'.

In contrast to Czech grammars (e.g., [13, pp. 221–222]), we do not distinguish interrogative pronouns as a separate subtype because of their unclear distinction from relative pronouns.

5.3 Possessivity and reflexivity

Several subtypes of the pronouns are introduced based on the feature of reflexivity and possession. These features are characteristic of particular personal and relative pronouns.

Besides gender and number of an object, **possessive pronouns** (for the 3rd person) also express the number and gender of the possessor; e.g., *jeho chalupy* 'his cottages' (fem. pl., possessor: masc. sg.), *z jejíhož domu* 'from whose house' (masc. sg., possessor: fem. sg.). This is why the 6th (possgender) and the 7th (possnumber) tag positions are also filled; see the comparison with the other possessive pronouns which express only the number of a possessor (e.g. *můj dům* 'my house' (possesor: sg.), *náš dům* 'our house' (possesor: pl.)) or nothing (e.g, *hájí svoji pravdu* 'defends his/their truth') in Tab. 9.

Reflexive pronouns express only a limited number of morphological categories. Possessive reflexive pronouns only express agreement in gender, number, and case. Gender and number of the possessor are not distinguished. Personal reflexive pronouns express only case (e.g., *mluví o sobě* (Loc.) 'he talks about himself'; *zatleskejte si* (Dat.) 'give yourselves a clap'), cf. Tab. 9.

Form	Lemma	Tag
můj dům	můj	`PSYS1-S1-------`
náš dům	náš	`PSYS1-P1-------`
jeho chalupy	jeho	`P9XXXZS3-------`
z *jejíhož* domu	jehož	`P1ZS2FS3-------`

Form	Lemma	Tag
hájí *svoji* pravdu	svůj	P8FS4----------
mluví o *sobě*	se	P6--3----------
zatleskejte *si*	se	P7--3----------

Tab. 9. Examples of possessive and reflexive pronouns

5.4 Clitichood

Some of the personal pronouns have a very wide set of forms. In addition to the basic variants, they often have clitic (weak) variants used in specific syntactic contexts (e.g., *zná ho dobře* 'he knows him well'; *dej mi to* 'give me that', and *zatleskejte si* 'give yourselves a clap') which have a special value of the 2nd tag position; cf. Tab. 9 and 10. Distinguishing clitic forms at the 2nd tag position violates the principle that the 2nd position is the same for the whole paradigm [7, p. 7]. This exception follows from the historical development of the MorfFlex dictionary.

Form	Lemma	Tag
zná jen *jeho*	on	PEYS4--3-------
zná *ho* dobře	on	P5ZS4--3-------
pro *něho*	on	PEZS4--3------1
proň	on	PEZS4--3-----p-
dej *mi* to	já	PH-S3--1-------
dej *mně* to	já	PP-S3--1-------

Tab. 10. Examples of personal pronouns

Furthermore, several pronouns (e.g., *on* 'he', *jenž* 'who/what/which') have a special form when following a preposition (cf. *zná jen jeho* (Accus. sg.) 'he knows only him' vs. *pro něho* (Accus. sg.) 'for him', or its rarely used form *proň*[6]). These forms are marked at the 15th position of the tag (cf. Tab. 10).

6 CONCLUSION

Even in Czech linguistics, there are many approaches to these POS types, each with its pros and cons. We have proposed one that primarily takes into account the morphological aspect. In the multi-layer concept of language [15] applied to PDT-C, syntactic and semantic properties are captured in higher layers. In the proposal under

[6] The form *proň* belongs to the so called "aggregates", forms created by joining two or more forms. Note the letter *p* (for the preposition *pro*) at the 14th tag position (in Tab. 10). For more details see [14].

discussion, numerals and pronouns are sorted at the 2nd tag position according to their morphological behavior combined with the traditional semantic classification. The value of the 2nd tag position determines the type of declension (adjectival, nominal, etc.) or the uninflected character of word, as well as what other morphological categories are expressed in the given subtype.

ACKNOWLEDGEMENTS

The research and language resource work reported in the paper has been supported by the LINDAT/CLARIAH-CZ project funded by Ministry of Education, Youth and Sports of the Czech Republic (LM2018101).

References

[1] Hajič, J. et al. (2020). Prague Dependency Treebank – Consolidated 1.0 (PDT-C 1.0). LINDAT/CLARIAH-CZ, Prague. Accessible at: http://hdl.handle.net/11234/1-3185.

[2] Hajič, J., Hlaváčová, J., Mikulová, M., Straka, M., and Štěpánková, B. (2020). MorfFlex CZ 2.0. LINDAT/CLARIAH-CZ, Prague. Accessible at: http: //hdl.handle.net/11234/1-3186.

[3] Petkevič, V., Hlaváčová, J., Osolsobě, K., Svášek, M., and Šimandl, J. (2019). Parts of Speech in NovaMorf, a New Morphological Annotation of Czech. Jazykovedný časopis 70(2), pages 358–369.

[4] Komárek, M. et al. (1986). Mluvnice češtiny 2. Academia, Prague.

[5] Štícha, F. et al. (2018). Velká akademická gramatika spisovné češtiny. Academia, Prague.

[6] Hajič. J. (2004). Disambiguation of Rich Inflection (Computational Morphology of Czech). Karolinum, Prague.

[7] Mikulová, M. et al. (2020). Manual for Morphological Annotation, Revision for the Prague Dependency Treebank – Consolidated 2020 release. Technical report, 2020/TR-2020–64, Charles University, Prague.

[8] Slovenský národný korpus – prim-6.1-public-sane. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2013. Accessible at: https://korpus.juls.savba.sk/.

[9] Morphological annotation of texts in the Slovak National Corpus – Numerals and Pronouns. Accesible at: https://korpus.sk/num.html; https://korpus.sk/pronom.html.

[10] https://www.sketchengine.eu/polish-nkjp-part-of-speech-tagset/.

[11] Prepiórkovski, A. (2009). A comparison of two morphosyntactic tagsets of Polish. In Representing Semantics in Digital Lexicography: Proceedings of MONDILEX Fourth Open Workshop, Warsaw, pages 138–144.

[12] Petkevič, V. (2010). L'accord en tchèque: le centre et la périphérie. Écho des études romanes, 6(1–2), pages 143–160.

[13] Cvrček, V. et al. (2015). Mluvnice současné češtiny. Karolinum, Prague.

[14] Hlaváčová, J., Mikulová, M., Štěpánková, B., and Hajič, J. (2019). Modifications of the Czech morphological dictionary for consistent corpus annotation. Jazykovedný časopis 70(2), pages 380–389.

[15] Sgall, P., Hajičová, E., and Panevová, J. (1986). The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Academia, Prague.

sciendo
DOI 10.2478/jazcas-2021-0043

ENGLISH DETACHED ADJECTIVAL CONSTRUCTIONS WITH AN EXPLICIT SUBJECT: A QUANTITATIVE CORPUS-BASED ANALYSIS

VIKTORIIA ZHUKOVSKA
Kyiv National Linguistic University, Kyiv, Ukraine

ZHUKOVSKA, Viktoriia: English detached adjectival constructions with an explicit subject: A quantitative corpus-based analysis. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 465 – 474.

Abstract: This article reports on the quantitative corpus-based investigation into the form-function interplay of the English detached adjectival construction with an explicit subject. Taking Usage-based Construction Grammar as its theoretical framework, this paper investigates the patterns of attraction of lexical items that appear in the main slots of the grammatical construction. The data obtained substantiate the constructional status of the construction and determine its semantic and functional specification in present-day English.

Keywords: detached clauses, Usage-based Construction Grammar, grammatical construction, simple collexeme analysis

1 INTRODUCTION

The English detached adjectival clauses with an explicit subject can be exemplified by the following sentences taken from the BNC-BYU corpus [1]:

1) *Tsu Ma looked up, tears filling his eyes,* ***[his voice soft]****.*
2) ***[Her glass empty]****, she accepted another from Lucenzo.*

The syntactic pattern under study represents adjectival secondary predication of syntactically independent configuration. It is part of a minimally bi-clausal structure consisting of a matrix clause and an adjectival clause with its own explicitly expressed subject, separated from the matrix by a punctuation mark. The syntactic pattern has a fixed binary structure [NP XP], where (NP) is a secondary subject, distinct from the subject of the matrix clause, and (XP) is a predicative group with an adjective phrase (AdjP) as a secondary predicate. The pattern can be attached to the matrix clause through augmentors (mostly *with*) or asyndetically. In a sentence, the pattern performs the general syntactic function of an adverbial modifier elaborating, extending, or enhancing the matrix proposition. Regarding the form, the obligatory slots of the pattern are schematically represented as [øaug/aug][Subj][PredAdjP].

Although a considerable amount of research has been devoted to the English non-finite clauses with an explicit subject ([2], [3], [4] to name but a few), no study, however, has so far dealt comprehensively with the semantic and functional properties of the detached verbless clauses, especially of the adjectival type, by

conducting a quantitative corpus-based investigation. Based on empirical data drawn from the BNC-BYU, this study focuses on the form-function interplay of the analysed pattern to gain information about its constructional status and idiosyncratic semantic and functional features in modern English usage.

2 THEORETICAL AND METHODOLOGICAL ASSUMPTIONS

2.1 Theoretical background

In our study, we follow the theoretical and methodological premises of Usage-based Construction Grammar ([5], [6], [7]). This cognitive linguistic theory offers a comprehensive way of analysing both general and idiosyncratic properties of language units and recognizes frequency of occurrence as a factor influencing the degree of their entrenchment in a speech community [8].

From the construction grammar perspective, we take for granted that English detached adjectival clauses with an explicit subject are *constructions* since they instantiate sufficiently frequent pairings of form and content (meaning/function). As a clausal type of constructions, the pattern elaborates the meaning in a way of discourse functions rather than coded semantics:

FORM: [øaug/aug][SubjNP][PredNF/VL]] ↔ MEANING: […]FUNCTION

The construction represents a node in a taxonomic constructional network of English detached non-finite and verbless constructions with an explicit subject. The network is organized around a constructional scheme, represented by a construction of the highest level of schematicity and abstractness – *macro-construction* (*dtcht-SubjPred*NF/VL–*cxn*).

In this study, we focus on the unaugmented (*øaug*) construction with AdjP as a predicate (*dtcht-øaugSubjPredAdjP–cxn*), based on the constructs collected from the BNC-BYU corpus. Adopting the usage-based perspective and applying the quantitative method of collostructional analysis, we discuss distributional properties of adjectives and nouns in the [Subj] and [Pred] slots of the construction as the parameter reflecting functional dynamics of the syntactic pattern.

2.2 Corpus, data and statistical procedure

The analysis of the *dtcht-øaug-SubjPredAdjP–cxn* is based on authentic English usage-data drawn from the well-balanced British National Corpus [1] in December 2020. The data were retrieved automatically using the BNC-BYU's search engine. In total, the queries yielded 857 tokens that were then checked manually to avoid spurious hits and formally similar but functionally different constructions (e.g., *Stir the tomatoes, **tomato pure**, wine and seasoning and bring to a boil*; *"I'm sorry about your Mandy, Pat, **heart sorry**. We all are."*). False hits being removed, the database included 376 tokens to analyse.

The method utilized for quantitatively processing the data is taken from the family of collostructional analysis developed by St. Stefanowitsch and A. Gries ([9], [10]). The collostructional analysis is a set of quantitative procedures (the simple collexeme analysis, the distinctive collexeme analysis, and the co-varying collexeme analysis), aimed at investigating how strongly lexemes are attracted to particular slots in a construction. Specifically, the simple collexeme analysis detects the collostructional preferences of a particular constructional slot and helps to elaborate the meaning of the construction. The method rests on the principle of semantic compatibility, i.e., "a word may occur in a construction if it is semantically compatible with the meaning of the construction" [10, p. 213].

To begin with, we applied the simple collexeme analysis to identify adjectives that are significantly more frequent in the slot [Pred] of the construction since this seemed to be the lexically more prominent, and hence linguistically more relevant slot. The collexeme analysis of the predicate slot was further supported by the output of the collexeme analysis of the nouns in slot [Subj]. The noun collexemes were analysed for their contribution to a more precise semantic and categorial specification of the adjective collexemes.

The calculations were performed using Coll.analysis 3.2a for R script [11]. The script adopts a Fisher-Yeats Exact test to identify significant collocational patterns and therefore yields reliable results even in cases of low-frequency tokens and is considered one of the most precise collocational tests [9].

3 SEMANTIC ANALYSIS

The database of this study consists of 376 tokens. As it turns out, the construction is of the highest frequency of occurrence among other types of English verbless detached constructions with an explicit subject. This construction appears with a frequency of 3.75 per million words, making it at best a mildly frequent pattern in English (see Table 1).

Construction	PredAdjP	PredPP	PredAdjP	PredNP
dtcht-unaug-SubjPredVL–cxn	3.75	3.04	0.57	0.54
dtcht-with-aug-SubjPredVL–cxn	2.34	3.51	1.87	0.12
dtcht-despite-aug-SubjPredVL–cxn	0.08	0.01	0.1	0.04
dtcht-without-aug-SubjPredVL–cxn	0.01	0.05	0.06	–
dtcht-what with-aug-SubjPredVL–cxn	0.01	0.01	0.01	–
Total	6.19	6.62	2.61	0.7

Tab.1. Overview of the normalized frequencies of the tokens in the BNC-BYU

Out of 151 adjective types, 89 items are used merely once with the pattern. They account for 58.94% of the total number of items in the construction. Lexemes

with low frequency are rather loosely associated with the pattern under study. At the same time, hapax legomena, i.e., items with a token frequency 1, define a potential productivity of the pattern. A bigger productivity ratio proves a higher potential productivity of the syntactic pattern and means that a greater number of new types will be produced based on the given constructional schema [12, p. 128]. The estimated productivity ratio of the analysed construction is not high (0.24) and signifies the pattern is of medium productivity in present-day English.

The token numbers suggest that at a lower level of abstraction the *dtcht-øaug-SubjPredAdjP* construction subsumes some adjective-specific constructions, such as *dtcht-øaug-SubjPred**wide**–cxn*, *dtcht-øaug-SubjPred**outstretched**–cxn*, etc., and a number of adjective-group specific constructions, such as *dtcht-øaug-SubjPredAdjP* (***DIMENSION/ PHYSICAL PROPERTIES/ SPEED/ COLOUR***)*–cxn* on a higher level of schematicity.

The collexeme analysis allows us to determine the semantic restrictions the construction imposes on the lexical items filling its main slots. The results of the analysis show that out of 151 adjective lexemes in the construction, 107 items reveal a significant attraction to the pattern (coll. strength > 1.30103 = $p<0.05$) and 10 adjectives are repelled from it. It should be highlighted, that the lower the p-value, the greater the probability that the observed frequency distribution of adjectives is not random, and the greater the attraction between the lexeme and the construction. The data suggest that only 49 lexemes reach the highest significance level (coll. strength > 3 = $p<0.001$). The highest scores indicate that these tokens most typically fill the slot [Pred] of the construction. Table 2 illustrates the first 10 attracted collexemes ranked according to the value of the collostruction strength.

	Adjectives	Coll.strength
1.	outstretched	114.64
2.	narrowed	86.66
3.	clenched	77.99
4.	closed	68.48
5.	wide	57.93
6.	bright	55.02
7.	flushed	52.94
8.	parted	46.42
9.	expressionless	37.10
10.	pale	34.87

Tab. 2. The top 10 significantly attracted adjective collexemes

The adjective collexemes are revealing of the semantic specificity of the analysed pattern. The meaning of the construction's collexemes is best comprehended on the basis of semantic frames, i.e., schematic knowledge structures that provide

important background knowledge of different types of events, relations or entities and participants in them [13], that were retrieved from the FrameNet project. As a frame element, an adjective is typically associated with the participant role Attribute. Within the 49 adjective collexemes of the construction the following semantic frames show up (presented in the order of collocation strength of adjectives most strongly attracted to the construction).

The first set of adjectives is constituted by the lexemes *outstretched* (rank 1), *narrowed* (2), *clenched* (3), *closed* (4), *parted* (8), etc. This set of items (10 lexemes) can be understood with reference to the *Body_part_posture* frame specifying what position or orientation the body or part of the body is in. This group predominantly includes lexemes of *V-ed* form derived from the respective verbal bases (*narrow*, *clench*, *close*), except for **outstretch*.

The meaning of the adjectives *bright* (6), *pale* (10), *ablaze* (15), *colourless* (43), etc. is understood within the *Colour_qualities* frame (6 items) that contains words nominating specific degrees of colour.

The set of adjectives *wide* (5), *deep* (36), *huge* (42) evoke the semantic fame *Dimension*, concerning words that express an object's measurement with respect to some attribute.

The most numerous group of adjective collexemes is constituted by the lexemes denoting a particular gradable attribute (*full* (13), *stiff* (16), *dry* (20), *hard* (29), etc.) (11 items). This set of adjectives evokes the frame *Measurable_attributes*, that describes an entity with a particular scalar attribute.

Another group of collexemes is constituted by the adjectives (7 items) *expressionless* (9), *impassive* (24), *grim* (19), *wild* (27), *angry* (36), etc., whose meaning can be interpreted regarding the *Emotions* frame. This frame specifies a particular emotional state of the experiencer, that may be indicted to an external observer by a body part or gesture.

The next category of strongly attracted adjectives (7 items) includes lexemes such as *untouched* (17), *unbrushed* (46), *unkempt* (47), *bloodshot* (38), etc. These adjectives describing salient parts of a human body instantiate a schematic knowledge structure *Body_description_part*.

The adjectives *husky* (14), *hoarse* (18), *harsh* (23), *muted* (45) are understood within the semantic frame *Sound_level* that describes entities judged by some sound level attribute.

As any semantic classification, the inclusion of adjectives into a semantic frame is not exclusive, an adjective may be attested to more than one frame because it is employed in more than one way. To maximize the precision of the semantic analysis of the adjectives associated with the slot [Pred] we carried out a collexeme analysis of common nouns in the slot [Subj] of the pattern. The analysis is expected to detect whether there are any constraints to be found on the construction's subject referents. Table 3 presents the top 10 out of 28 significantly attracted noun collexemes ($>3= p < 0.001$).

	Nouns	Coll.strength
1.	eye	213.31
2.	face	71.33
3.	voice	44.99
4.	arm	44.31
5.	mouth	40.30
6.	expression	26.83
7.	hand	21.86
8.	tone	20.29
9.	fists	18.89
10.	cheek	18.00

Tab. 3. The top 10 significantly attracted noun collexemes

The output of the collexeme analysis proves that the construction is highly restrictive regarding nouns in its subject position. All strongly attracted nouns evoke the *Body_parts* frame that contains somatisms, i.e., nouns naming limbs and their parts (*arms*, *hands*, *fists*, *legs*), external parts of the body (*face*, *ears*), and their constituent parts (*eyes*, *mouth*, *cheeks*, *lips*) or features (*expression*, *features*), and other elements of the human anatomy (*voice*, *gaze*, *tone*, *breath*). Thus, the subject slot of the *dtcht-øaug-SubjPredAdjP* construction is typically filled with inanimate non-volitional nouns. The referent of the construction's subject appears to be partially coreferent with the referent of the matrix subject (95% of all tokens), instantiating meronymic (whole-part) relations. Being exclusively modified by possessive pronouns *his*, *her*, *their*, *its*, *my*, *our*, *your*, the construction's subject referents nominate unalienable entities, namely parts of the body, of the matrix subject referent.

It becomes evident that the strongly associated adjectives functioning as predicates of *Body_part* subjects also reflect corporeal semantics, referring to the properties of a human being, describing and expressing physical characteristics, dimensions or position of a body and body parts, denoting human emotions and feelings. They are typically stage-level adjectives that render temporary properties of the subject referent. It does not mean that individual-level adjectives are impossible in the pattern. In the corpus sample, there are instances of adjectives that denote long-standing features of an entity. When attracted to the construction, these lexemes reveal statistically insignificant collostruction strength (e.g., *black* (coll. strength = 0.94), *obvious* (0.72), *blue* (0.60)), otherwise they are repelled by the pattern. The 10 repelled adjectives are *good*, *long*, *important*, *big*, *white*, *clear*, *easy*, *happy*, *serious,* and *dead*. One of the possible explanations why these adjectives are not common in the predicate slot of the construction can be their descriptive semantics that conditions their preferable occurrence in the prenominal position.

The results of the simple collexeme analysis carried out separately for adjectives and nouns occurring in the *dtcht-øaug-SubjPredAdjP* construction reveal that the pattern exhibits distinct semantic preferences for the lexemes in its subject and predicate slots.

4 FUNCTIONAL SPECIFICATION

The functional specification of the *dtcht-øaug-SubjPredAdjP* construction is primarily determined by the lexemes filling its [PredAdjP] slot since this slot seems to be the most informative. The pattern attracts adjectival lexemes of two types, adjectives genuine (*wide*, *open*, *pale*, *husky*, etc.) and adjectival past participles (*narrowed*, *clenched*, *curved*, *untouched*, etc.).

Syntactically, adjectives are used in two types of context: as pre- and postnominal adjectives (non-predicative and predicative, respectively). Predicative adjectives reflect temporary states or specific events while non-predicative adjectives express semantically permanent or characteristic features of the noun they modify [14, p. 81]. The adjective collexemes of the analysed construction represent a predicative type. Occurring in the slot [Pred] the stage-level adjectives ascribe a temporary or stage-like state to the subject referent. The whole construction acquires stative reading, where a state is rendered as holding for a while rather than being ascribed to the subject referent. Individual-level adjectives, though not statistically attracted but still not uncommon in the pattern, in predicate position might bleach their individual reading and acquire a more temporary character, coerced by the construction, as in (3).

3) *John of Gaunt looked up abruptly and stared like a hungry cat at Athelstan,* ***his eyes yellow, hard and unblinking.***

Past participles in English are analysed in terms of passives and subdivided into verbal passives and adjectival passives [15, p. 36]. Contrary to verbal passives, expressing canonical events where "an agent acts on a patient to induce a change of state" [16, p. 357], stative passives are qualified as unambiguously adjectival [3, p. 1440]. The verbal passive portrays the event as dynamic in which the entity is depicted as the Patient, while the stative passive construes the state of the entity resulting from the action denoted by the verb. This entity carries out the semantic role of the Theme, i.e., 'what is in a state or in a change of state' [5, p. 428].

The past participles in the predicate slot of the construction are "stative-adjectival" (*V-ed*) participles, generally profiling the final state of the process denoted by their verbal basis. Their adjectival status is confirmed by such diagnostic tests:

1) the absence of the verbal base of the participle and the use of the prefix *un-* activating the meaning of "the event that did not take place" [5, p. 427] (such as **untouch*, **unbrush*);

2) the unaccusative verbal base (*narrowed*, *muted*, *closed*);

3) the participles can be potentially modified by quantifiers (*more/most*, *too*, *very*).

Within the adjectival participles a specific *un-V-ed* type (*untouched*, *unbrushed)* should be discussed. This *un*-participle qualifies the state of the subject referent as 'not being exposed or subjected to V', i.e., a state due to the absence or non-occurrence of an action [5, p. 428].

The presented considerations are in accord with the usage-based construction grammar tenet of iconicity relations between a construction's form and meaning ([6], [7]). Occurring in the predicative position of the construction under scrutiny, the individual-level adjectives attribute some (temporary) property to the subject referent, while stage-level adjectives and adjectival participles (*V-ed* and *un-V-ed*) induce a stative reading and ascribe a state to the subject's referent. More specifically this state can be further qualified as a temporary state (construed by stage-level adjectives), a state resulting from an action (construed by *V-ed* adjectival past participles), and a state due to the absence an action (construed by *un-V-ed* adjectival past participles). Thus, the stage-level adjectives and adjectival participles in the pattern's predicate slot are deemed as subject-oriented depictives construing a property or state that holds of the entity during the event time of the matrix predicate.

The conducted analysis shows that the *dtcht-øaug-SubjPredAdjP* construction is not functionally homogeneous. We can identify two functions of the pattern: depictive and attributive, with the respectively construed properties and states of the subject referents. The attributive function is exemplified in (4).

4) *Too late -- Perdita,* ***her face ashen****, her black eyes blazing, had a pitchfork poised a foot from Raimundo's capacious buttocks*.

The construction elaborates on the matrix subject referent, specifying, describing or clarifying it through the exemplification of the property ascribed to its subject referent. The depictive function is represented by such instantiations as

5) *She gasped and stepped back,* ***her face pale****.*

6) *Her breasts heaving,* ***her throat dry****, she strained tensely to release herself.*

In these examples, the construction extends and enhances the main event construed by the matrix predicate by providing additional (new) details through the description of a (temporary/resultative/absent) state of its subject referent.

The depictive function can be considered prototypical due to its higher ratio in the analysed sample (354: 22). The attributive function is more peripheral, represented by a significantly lower number of its examples in the research database.

High collocational strength of nouns evoking *Body_parts* frame (*eyes* (rank 1), *face* (2), *voice* (3), etc.) can be attributed to the specific distribution of the *dtcht-øaug-SubjPredAdjP* construction in modern English usage. The corpus data suggest that the pattern is predominantly observed in the written discourse, especially in narrative/literary texts. The construction is exceptionally prominent in fiction (86.62% of all the tokens), where it serves as effective means of packing descriptive

information and providing additional details to the event in the matrix clause. Particularly in fiction somatisms provide information about the object they nominate and indirectly render various emotional, psychological, and physical properties or states of an individual [17, p. 3454].

With the prevalence of *Body_parts* nouns in the slot [Subj], only a part of the matrix event is profiled. The referents of the subject in the investigated construction expressed by inanimate nouns (parts of a human body) are construed as Themes of states rendered by the adjectival predicate of the pattern, with the Agent/Experiencer represented by the matrix subject.

5 CONCLUDING REMARKS

The results of the quantitative corpus-based analysis of the form-function interaction of the English *dtcht-øaug-SubjPredAdjP* construction suggest the following tentative conclusions.

The construction at hand instantiates adjectival secondary predication of syntactically independent configuration. This pattern is a mildly frequent construction, exhibiting medium productivity in present-day English.

The English *dtcht-øaug-SubjPredAdjP* construction displays a notable consistency in attracting nouns and adjectives of certain semantics to fill [Subj] and [Pred] slots. The quantitative corpus linguistic method of collostructional analysis has proved to be efficient for detecting highly attracted items revealing of the lexical preferences of the construction.

The investigated construction is linked with two functions. The instances of the constructions where the predicates ascribe properties to their subject referents, construing them as carriers of properties are indicative of the pattern's attributive function. The instances where the predicates ascribe a state to their subject referents, construing them as entities in a (temporal/resultative/absent) state represent the pattern's depictive function. The depictive function is viewed as prototypical, while the attributive function is more peripheral.

The functional specification of the analysed construction is conditioned by its register distribution. The syntactic pattern predominates in narrative/literary texts and utterly prevails in fiction, where it serves as a means of rendering information about the properties and states of the matrix subject referents. The subject referents denoting body parts express inalienable property, representing partially coreferential relations with the matrix subject referents. Being predominantly modified by possessive pronouns, the constriction's subject referents manifest pertinence relations with the subject referents of the matrix clause.

This study is of a preliminary character since the findings are obtained on the limited research material. Further more extensive corpus-quantitative research of the unaugmented construction and constructions introduced by the augmentors *with*,

without, *despite*, *what with* would be needed to achieve more reliability and corroborate the data received.

References

[1] BNC-BYU. Accessible at: https://www.english-corpora.org/bnc/.

[2] Kortmann, B. (1991). Free adjuncts and absolutes in English: Problems of control and interpretation. London, New York: Routledge, 253 p.

[3] Huddelson, R. D., and Pullum, G. K. (2002). The Cambridge Grammar of the English language. Cambridge: Cambridge University Press, 1842 p.

[4] Hasselgård, H. (2012). Possessive absolutes in English and Norwegian. In C. Fabricius-Hansen and D. T. T. Haug (eds.), Big events, small clauses: The grammar of elaboration, pages 229–258, Berlin, Boston: De Gruyter.

[5] Schönefeld, D. (2015). A constructional analysis of English un-participle constructions. Cognitive linguistics, 26(3), pages 423–466.

[6] Hoffmann, T. (2019). The more data, the better: A usage-based account of the English comparative correlative construction. Cognitive Linguistics, 30(1), pages 1–36.

[7] Horsch, J. (2020). Slovak comparative correlative CC' constructions from a construction grammar perspective. Jazykovedný časopis, 71(1), pages 25–40.

[8] Hilpert, M., and Diessel, H. (2017). Entrenchment in Construction grammar. In H.-J. Schmid (ed.), Language and the human lifespan series. Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge, pages 57–74. American Psychological Association; De Gruyter Mouton.

[9] Gries, S. T. (2015). The role of quantitative methods in cognitive linguistics: corpus and experimental data on (relative) frequency and contingency of words and constructions. In J. Daems, E. Zenner, K. Heylen, D. Speelman and H. Cuyckens (eds.), Change of paradigms – new paradoxes: Recontextualizing language and linguistics, pages 311–325. Berlin & New York: De Gruyter Mouton.

[10] Stefanowitsch, A., and Gries, S. T. (2003). Collostructions: Investigating the interaction between words and constructions. In International Journal of Corpus Linguistics 8(2), pages 209–243.

[11] Gries, S. T. (2007). Coll.analysis 3.2a. A program for R for Windows 2.x. Accessible at: http://www.linguistics.ucsb.edu/faculty/stgries/teaching/groningen/index.html.

[12] Hilpert, M. (2013). Constructional change in English. Developments in allomorphy, word formation, and syntax. Cambridge: Cambridge University press, 233 p.

[13] FrameNet. Accessible at: https://framenet.icsi.berkeley.edu.

[14] Demonte, V. (2008). Meaning-form correlations and adjective position in Spanish. In L. McNally and C. Kennedy (eds.), Adjectives and adverbs. Syntax, semantics, and discourse, pages 71–100, New York: Oxford University Press.

[15] Dryer, M. S. (1985). The Role of thematic relations in adjectival passives. Linguistic Inquiry, 16(2), pages 320–326.

[16] Langacker, R. W. (2008). Cognitive grammar. A Basic Introduction. Oxford: Oxford University Press, 561 p.

[17] Frith, C. (2009). Role of facial expressions in social interactions. In Philosophical Transactions of the Royal Society, № 364, pages 3453–3458.

NATURAL LANGUAGE PROCESSING AND CORPUS BUILDING

sciendo
DOI 10.2478/jazcas-2021-0044

USING A PARALLEL CORPUS TO ADAPT THE FLESCH READING EASE FORMULA TO CZECH

KLÁRA BENDOVÁ
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

BENDOVÁ, Klára: Using a parallel corpus to adapt the Flesch Reading Ease formula to Czech. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 477 – 487.

Abstract: Text readability metrics assess how much effort a reader must put into comprehending a given text. They are, e.g., used to choose appropriate readings for different student proficiency levels, or to make sure that crucial information is efficiently conveyed (e.g., in an emergency). Flesch Reading Ease is such a globally used formula that it is even integrated into the MS Word Processor. However, its constants are language-dependent. The original formula was created for English. So far it has been adapted to several European languages, Bangla, and Hindi. This paper describes the Czech adaptation, with the language-dependent constants optimized by a machine-learning algorithm working on parallel corpora of Czech and English, Russian, Italian, and French, respectively.

Keywords: complexity, parallel corpus, Czech, Flesch Reading Ease, machine learning

1 INTRODUCTION

This study describes a machine learning-based approach to adapting the widely known Flesch Reading Ease [1] formula to Czech, based on a parallel corpus [2].

A written text is always a message conveyed by the author to the recipient without real-time interaction. Therefore, the author must assess the intended reader well, regarding their knowledge of the topic and contexts, but also their reading comprehension skills. This is immensely important, whenever lives and health, security, democracy, or property are at stake.

This is where the concept of readability comes into play. DuBay [3, p. 6] has summarized its most prominent definitions: „readability is the ease of reading created by the choice of content, style, design, and organization that fit the prior knowledge, reading skill, interest, and motivation of the audience“. Particularly in the English-speaking community, quantitative assessment of readability has been worked on since the early 20th century. The 1980's have already seen over 200 different readability formulas, with over a thousand studies attesting to their strong theoretical and statistical validity [3].

One of the most common readability formulas is Flesch Reading Ease [1], which is even implemented in the MS Word editor. On a scale from 0 to 100, it

measures the „ease“ of the text, using general features such as average length of sentences in words and average length of words in syllables, and a few constants. However, these constants are language-dependent. Šlerka and Smolík [4] found out in their pilot experiment that the Flesch Reading Ease was associated with the intuitively perceived linguistic complexity of different text genres even in Czech. However, the scores would not fall between 0 and 100. Due to inflections and the absence of articles, both making the average Czech word longer than English, any natural Czech text would score as difficult. Even common newspaper texts reach negative values, beyond the extreme difficulty end of the English scale.

This study will (1) introduce a selection of tools assessing diverse complexity features of Czech and other, mainly Slavic, languages; (2) describe the Flesh Reading Ease formula and its existing language adaptations; (3) describe the data and its pre-processing to derive the Czech parameters for Flesch Reading Ease; (4) describe the experiment; (5) report and interpret its results. Its goal is to offer a Czech-tailored replacement for the original English-based Flesch Reading Ease for the assessment of the readability of Czech texts.

2 RELATED WORK

2.1 Tools

There are numerous online tools to assess readability of English texts by diverse formulas. Nevertheless, this section will only list tools that were immediately relevant for this study. These are mostly tools tailored to Czech or Polish, and a multilingual tool that is still in development.

One of the most inspiring tools is EVALD [5], which primarily assesses text cohesion and coherence in pupils' essays, predicting the grade given by teachers. It is partly based on international readability formulas Flesh Reading Ease [1], Flesch-Kincaid Grade Level Formula [6], Coleman-Liau index [7], SMOG index [8], but none of them has been adapted to Czech. Apart from the cohesion/coherence assessment, EVALD has also been trained on Czech texts written by foreigners to guess the CEFR [9] proficiency level of their authors [10].

Another text assessment tool for Czech is QuitaUp [11], which mainly captures stylometric characteristics such as TTR, h-point, entropy, or word distance.

However, neither is a dedicated readability tool like e.g., the Polish Jasnopis [12], which combines statistical features with empirically measured reading comprehension.

Eventually, a multilingual readability assessment platform is in development (Common Text Analysis Platform – CTAP) [13]. It aggregates 600 textual features ranging from syllable count to lexical sophistication tailored to the languages currently represented: English and German. Other languages being worked on are Italian, French, Portuguese, Greek, and Czech.

3 FLESCH READING EASE

3.1 The original English formula

Flesch Reading Ease, presented by Rudolf Flesch [1], is defined as follows:

FRE = 206.835 – 84.6 wl – 1,015 sl[1],

where FRE = Flesch Reading Ease
wl = average word length in syllables
sl = average sentence length in words.

Henceforth I will refer to exact values as coefficients, while calling the variable formula elements parameters.

The results of Flesch Reading Ease virtually always fit within the range of 0 – 100. The higher the score, the higher the "ease," that is, the more its complexity decreases, and the lower education is expected in the reader to be well equipped to comprehend the text.

Flesch interprets these results in the book The Art of Readable Writing [14] as follows:

Reading Ease Score	Style Description	Estimated Reading Grade
0 to 30:	Very Difficult	College graduate
30 to 40:	Difficult	13th to 16th grade
50 to 60:	Fairly Difficult	10th to 12th grade
60 to 70:	Standard	8th and 9th grade
70 to 80:	Fairly Easy	7th grade
80 to 90:	Easy	6th grade
90 to 100:	Very Easy	5th grade

Tab. 1. Flesch Reading Ease Index interpretation

When computing this formula, Flesch was drawing on a formula he had invented in 1943. He skipped affix counts since they had proved troublesome to count for the formula users. Instead, he transformed this feature into syllable count, which he considered more mechanical and thus less error-prone [15]. However, Flesch used the omitted counts to determine the coefficients.

[1] I am quoting the paper *A New Readability Yardstick* [1] from DuBay's compilation of readability studies *Unlocking Language: The Classic Readability Studies* [3], where the decimal separator is misplaced (FRE = 206.835 – 84.6 wl – 1.015 sl), whereas the formula correctly reads FRE = 206.835 – 84.6 wl – 1,015 sl.

3.2 Language mutations of Flesch Reading Ease

Even formulas that use very generic features are as heavily language-dependent, as languages differ with respect to their phonological, morphological, and syntactic features. Individual languages need individual formulas. This also even applies to Flesch Reading Ease. Guryanov et al. [16] interpret its parameters as follows: WL (word length as the ratio between total syllables and total tokens) renders the information load of the text; short words make the text less informative than long words. SL (sentence length as the ratio between total words and total sentences) reflects cohesion; that is, cohesion decreases with the sentence length. This difference is language-dependent. I. V. Oborneva [17] observed that an average English word has 2.97 syllables, while an average Russian word has 3.29 syllables. This necessarily affects the coefficients; the more so if the results are supposed to span the same scale and be cross-linguistically comparable.

Currently there are formulas for Italian, French, Spanish [18], German [19], Russian [17], and Danish, as well as for Bangla and Hindi [20]. Garais [18] also mentions a Japanese formula, but the source is not sufficiently quoted.

The formulas were designed at different times, with different methods available then. The more recent formulas draw on machine-learning algorithms run over large data, including parallel corpora, while older formulas are based on sophisticated calculations.

For French, the first Flesch versions were calculated in 1958 [21] and 1963 [22], to be replaced by a third version [23], which is still in use [24]. Despite extensive research, unfortunately limited to the English-written literature, I have failed to find this current version for French and had to resort to the 1958 version [21].

FRE(French) = 207 – 1.015(total words/total sentences) – 73.6(total syllables/ total words)

The first version of the Russian formula was designed by Matkovskij in the 1970's. Matkovskij grounded his formula in the fact that Russian words have, on average, more syllables than English words and, therefore, he replaced one of the parameters with the number of tokens that have more than three syllables (Matkovskij, 1976 in [25]).

FRE(Russian_mod)= 0.62(total words/total sentences) + 0.123 X3 + 0,051

where X3 = the percentage of tokens with more than three syllables.

A more recent Russian version came from Oborneva in 2006. As already mentioned above, Oborneva based her calculations on the difference in number of syllables in Russian and English words [16], drawing on *Slovar russkogo yazyka pod redaktsyey Ozhegova* (39174 words) and *Muller English-Russian dictionary* (41977 words). In addition, she analyzed six million words of parallel Russian-English literary texts [16], her work resulting in the following formula [17]:

FRE(Russian) = 206.835 – 1.3(total words/total sentences) – 60.1(total syllables/total words).

4 DATA

The experiment is based on a cross-lingual comparison of parallel texts; therefore I used data from the InterCorp parallel corpus ([26], [2]). This corpus has Czech as the pivot language: all texts have a Czech version, which is manually sentence-aligned with at least one different language. Foreign languages are never directly aligned with each other, but through Czech. The Czech texts are both original texts, as well as translations. Among foreign texts, originals or translations from Czech were preferred during the acquisition, but translations from other languages are present as well. The corpus primarily comprises fiction, but also non-fiction and legal texts from the multilingual official production of the EU bodies. Tab. 2 shows the distribution of selected languages in InterCorp.

Language	**Czech**	**English**	**Russian**	**French**	**Italian**
Total of texts	586	348	128	233	136
Total of sentences	3 719 974	2 364 684	855 584	1 160 089	992 008
Total of tokens	43 446 132	33 190 659	9 449 802	18 921 311	14 466 499

Tab. 2. Distribution of the data used

5 METHOD

This section describes the actual experiment. Its goal was adaptation of Flesch Reading Ease to Czech and assessment of its validity by comparison with formulas for other languages.

The parameters of the Flesch Reading Ease formula are counts of words, syllables, and sentences. The InterCorp data came as XML files with tokenization and sentence splitting. I tested the sentence splitting with UDPipe [27], with no resulting corrections. I used the same method to count words and sentences in all languages.

On the other hand, syllable counting required individual language-specific scripts, since the phonotactic rules, as well as phoneme distributions, are language-specific. My syllable-counting scripts were based on a syllable-counting script by David Lukeš from the Institute of the Czech National Corpus, which considers the pitch (a vowel, diphthong, or a syllabic consonant), rather than syllable boundaries. Another option was using the PyHyphen library[2], as done for instance in Jasnopis [12], but my rule-based scripts were giving better results in manual sample checks. However, both approaches had problems counting syllables in French. The complexity of French

[2] Available at: https://pypi.org/project/PyHyphen/.

syllable counting can explain worse experiment results for French. When processing Russian, I considered only vowels to form syllables, drawing on [25].

Figure 1 shows the curves of the language specific FRE scores on parallel texts from InterCorp. Considering the English curve the reference, the Russian FLE fits it far better than the French and Italian. This implies that the Italian and French formulas, at least in my implementation, are less suitable to train the Czech formula adaptation than Russian, to achieve the best possible fit to English.

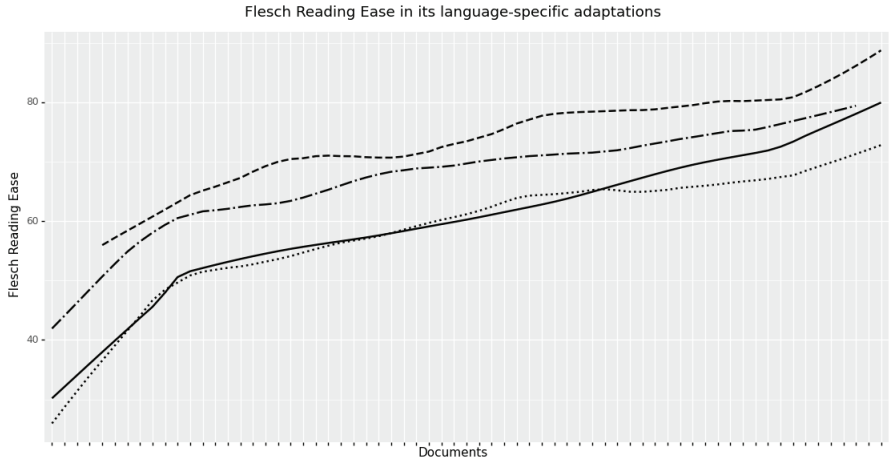


Fig. 1. Flesch Reading Ease in its language-specific adaptations. Solid line _ for English, dotted line **...** for Russian, dashdot line for French and dashdash line for Italian

To quantify the deviations seen in the plot in Fig. 1, I computed the RMSE (Root Mean Squared Error, a standard deviation evaluation in machine learning) of each language-specific FRE to the English FRE on English (Tab. 3). The French and Italian RMSE are indeed substantially higher, as expected based on Fig. 1.

	RMSE
English	–
Russian	5.100
French	10.518
Italian	12.991

Tab. 3. Root mean squared error for every language-dependent FRE used on Czech documents compared to English FRE used on the corresponding documents in English

The scatterplot in Figure 2 shows the FRE curves of Czech documents computed with the individual language-specific formulas; that is, the English, French, Italian, and Russian FRE for each Czech document, distinguished by the point shape. The solid line shows the *English* FRE of the corresponding *English*

versions of the Czech documents, as a reference of accuracy. The documents (on the X-axis) are ordered according to the English FRE of their English versions. There is an observable difference between the reference English curve and the curve representing the English FRE on the Czech documents. The original English FRE formula presents the Czech texts almost twenty points lower, which says that the Czech texts are two reading proficiency levels more difficult than their English versions.

Without fitting FRE to Czech, the best language-dependent formula to use would be the Russian one. It can certainly lie in the closeness of these two languages, but it can also be attributed to the worse fit of the French and Italian formulas to the English original (cf. Fig. 1).

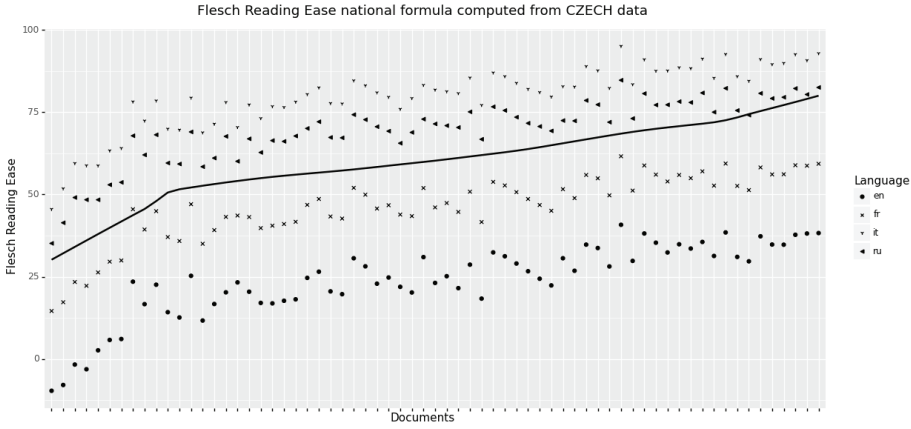


Fig. 2. Language-specific Flesch Reading Ease formula computed on Czech data. Solid line _ for English as the reference value

To find the optimal parameters for Flesch Reading Ease, I used the optimize. curve_fit algorithm from the SciPy library [28] with Russian and English separately. I neglected French and Italian due to their substantially worse fit to English. On input, the algorithm got FRE values of the individual Czech documents computed with the corresponding formula for the reference language. The algorithm compared these values with the values of the corresponding documents in the corresponding foreign language. The outcome was two different FRE functions for Czech.

I repeated the experiment with documents chunked into 100-sentence batches to increase the number of observations. The English and Russian inputs increased from 348 observations to 19,722 and 128 to 6,138 observations, respectively. However, this has not affected the best fit made on Russian texts, shown in Tab. 4. The best result was obtained using whole Russian texts as reference with RMSE 3.748 on test data.

Text types (number)	FRE for CZECH	RMSE test data
EN texts (347)	$206.835 - 1.424(\frac{\text{tot words}}{\text{tot sentences}}) - 63.920(\frac{\text{tot syllables}}{\text{tot words}})$	6.039
EN parts (19 722)	$206.835 - 1.672(\frac{\text{tot words}}{\text{tot sentences}}) - 62.182(\frac{\text{tot syllables}}{\text{tot words}})$	4.639
RU texts (127)	$\mathbf{206.835 - 1.388(\frac{\text{tot words}}{\text{tot sentences}}) - 65.090(\frac{\text{tot syllables}}{\text{tot words}})}$	**3.748**
RU parts (6 138)	$206.835 - 1.514(\frac{\text{tot words}}{\text{tot sentences}}) - 60.096(\frac{\text{tot syllables}}{\text{tot words}})$	4.363

Tab. 4. Version of Flesch Reading Ease for Czech language and the RMSE computed for test data

For the final evaluations, I merged the train and test data for English, Czech, and Russian, respectively, and computed the RMSE between the Czech FRE and English FRE, as well as the RMSE between the Russian and the English FRE. The Czech-English RMSE is 5.067, which is better than RMSE for Russian and English with 5.100 (Tab. 5).

	RMSE
English	–
Russian	5.100
French	10.518
Italian	12.991
Czech	**5.067**

Tab. 5. Root mean squared error for every language compared to English

Figure 3 confirms that the Czech language specific FRE on Czech texts is closest to the English FRE on English texts from all available language specific FREs on their languages.

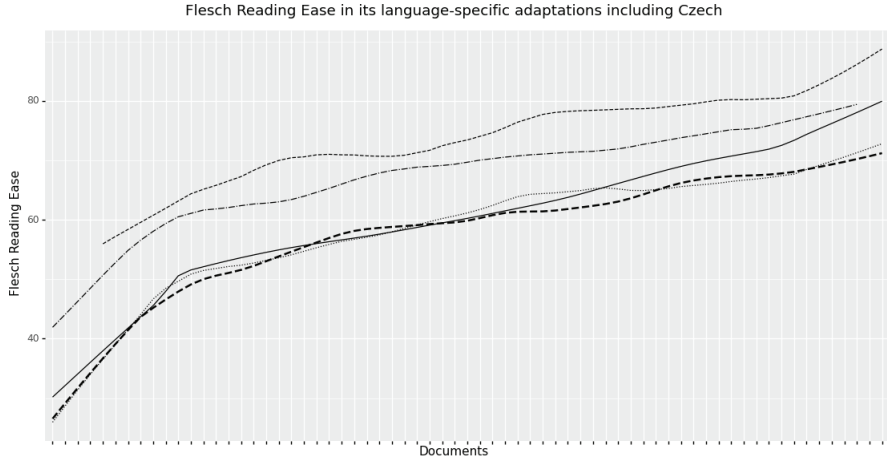


Fig. 3. Flesch Reading Ease in its language-specific adaptations including Czech. Original lines from Fig. 1 are thin, while the one for Czech formula -- is thick

6 DISCUSSION

Although the parallel corpus is relatively small and it is not balanced, the Czech formula superseded the (obsolete) French and Italian formulas. The relatively poor fit of the French and Italian formulas to English compared to Russian in this exercise can be blamed on possible conceptual errors in my syllable-counting scripts, while the fit on Russian was so much better because substantial syllable-conceptualization differences are very unlikely in this language pair. I have reached the maximum possible fit given the available data and language-dependent formulas.

This work is part of a larger project. The Czech adaptations of this and other readability formulas and features are to be implemented in CTAP [13]. The script is freely available at GitHub [29].

This entire approach naturally draws on the assumption that translations have the same readability as originals. Good translations are supposed to be semantically and stylistically faithful, as well as idiomatic. Given that InterCorp comprises mainly professionally published fiction and official multilingual documents, the translation quality is maintained.

The statistics (word and syllable counts) on which the current FRE is based are seemingly primitive, but Flesch himself proved them to strongly correlate with much more sophisticated statistics he had used earlier. In the original formula versions, Flesch made use of the contemporary psychological and pedagogical knowledge and found text features to reflect how "conversational", "personal", and "interesting" a text passage be, considering also text cohesion by counting pronouns, personal names, and nouns referring to humans. Besides, he accounted for the conceptual complexity (abstraction) by counts of lexical derivatives [15, p. 101]

These units are so essential for the content that they do not leave much room for deviation between languages. This suggests that, although their counts will be different in translation pairs (e.g., pronouns between a pro-drop and non-pro-drop language), their distributions within each language will be similar. The dissimilarity creates the documented error margins of the individual formula adaptations.

ACKNOWLEDGEMENTS

This work was supported by the Czech Science Foundation grant 19-19191S: Linguistic Factors of Readability in Czech Administrative and Educational Texts. The work described herein has also been using data/tools/services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

References

[1] Flesch, R. (1948). A New Readability Yardstick. Journal of Applied Psychology, 32, pages 221–233.

[2] Rosen, A. (2016). InterCorp – a look behind the façade of a parallel corpus. In Polskojęzyczne Korpusy Równoległe Polish-Language Parallel Corpora, pages 21–40, Instytut Lingwistyki Stosowanej, Warszawa.

[3] DuBay, W. (2007). Smart Language. Readers, Readability, and the Grading of Text. Impact Information, Costa Mesa, California.

[4] Šlerka, J., and Smolík, F. (2010). Automatická měřítka čitelnosti pro česky psané texty. Studie z Aplikované Lingvistiky, 1, pages 33–44.

[5] Novák, M., Mírovský, J., Rysová, K., Rysová, M., and Hajičová, E. (2019). EVALD 4.0 – Evaluator of Discourse. Accesible at: http://hdl.handle.net/11234/1-3065.

[6] Kincaid, J. P., Fishburne, R. P., Rogers, R. L., Chissom, B. S., and BRANCH, N.T.T.C.M.T.R. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula). for Navy Enlisted Personnel. Defense Technical Information Center. Accessible at: https://books.google.cz/books?id=7Z7ENwAACAAJ.

[7] Coleman, M., and Liau, T. L. (1975). A computer readability formula designed for machine scoring. Journal of Applied Psychology, 60, pages 283–284.

[8] McLaughlin, G. H. (1969). SMOG grading – a new readability formula. Journal of Reading, 22, pages 639–646.

[9] Council of Europe. (2018). Common European Framework of Reference for Languages: Learning, Teaching, Asessment. Companion volume. Council of Europe Publishing, Strasbourg. Accesible at: https://www.coe.int/lang-cefr.

[10] Rysová, K., Rysová, M., Mírovský, J., and Novák, M. (2017). Introducing EVALD – Software Applications for Automatic Evaluation of Discourse in Czech. RANLP Proceedings, Bulgaria, pages 634–641.

[11] Cvrček, V., Čech, R., and Kubát, M. (2020). QuitaUp. Czech National Corpus and University of Ostrava. Accesible at: https://www.korpus.cz/quitaup/.

[12] Dębowski, Ł., Broda, B., Nitoń, B., and Charzyńska, E. (2015). Jasnopis – A Program to Compute Readability of Texts in Polish Based on Psycholinguistic Research. Natural Language Processing and Cognitive Science, 2015 Libreria Editrice Cafoscarina, Venezia, Italy, pages 51–61.

[13] Chen, X., and Meurers, D. (2016). CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. Apollo – University of Cambridge Repository. Accesible at: https://www.repository.cam.ac.uk/handle/1810/292470.

[14] Flesch, R. (1974). The art of readable writing. 2nd ed. Harper, New York.

[15] DuBay, W. H. (2008). Unlocking Language: Classic Readability Studies. IEEE Transactions on Professional Communication, 51.

[16] Guryanov, I., Yarmakeev, I., Kiselnikov, A., and Harkova, I. (2017). Text Complexity: Periods of Study in Russian Linguistics. Revista Publicando, 4, pages 616–625.

[17] Oborneva, I. V. (2006). Mathematical model for evaluation of didactic texts. Proc of Moscow State Pedag Univ, 4, pages 141–147.

[18] Garais, E.-G. (2011). Web Applications Readability. Romanian Economic Business Review, 5, pages 117–121.

[19] Amstad, T. (1978). Wie verständlich sind unsere Zeitungen? Studenten-Schreib-Service.

[20] Sinha, M., Sharma, S., Dasgupta, T., and Anupam, B. (2012). New Readability Measures for Bangla and Hindi Texts.
[21] Kandel, L., and Moles, A. (1958). Application de l'indice de flesch à la langue française. Cahiers Etudes de Radio-Télévision, 19, pages 253–274.
[22] De Landsheere, G. (1963). Pour une application des tests de lisibilité de Flesch à la langue française. Le Travail Humain, pages 141–154.
[23] Henry, G. (1975). Comment mesurer la lisibilité. Labor, Brussels, Belgium.
[24] François, T., and Fairon, C. (2012). An AI readability formula for French as a foreign language, 477 p.
[25] Solnyshkina, M., Ivanov, V., and Solovyev, V. (2018). Readability Formula for Russian Texts: A Modified Version: 17th Mexican International Conference on Artificial Intelligence, MICAI 2018, Proceedings, Part II, pages 132–145.
[26] Čermák, F., and Rosen, A. (2012). The Case of InterCorp, a multilingual parallel corpus. International Journal of Corpus Linguistics, 13, pages 411–427.
[27] Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics, Brussels, Belgium, pages 197–207.
[28] SciPy 1.0 Contributors, Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T. et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods, 17, pages 261–272.
[29] https://github.com/vanickovak/ReadabilityFormula.

sciendo
DOI 10.2478/jazcas-2021-0045

A SYNCHRONIC AND DIACHRONIC COMPUTER CORPUS OF MAKARSKA LITTORAL DIALECTS (CROATIA)

JURAJ BENIĆ[1] – LOBEL FILIPIĆ[2]
[1] Faculty of Mechanical Engineering and Naval Architecture, University of Zagreb, Zagreb, Croatia
[2] Institute of Croatian Language and Linguistic, Zagreb, Croatia

BENIĆ, Juraj – FILIPIĆ, Lobel: A synchronic and diachronic computer corpus of Makarska littoral dialects (Croatia). Journal of Linguistics, 2021, Vol. 72, No 2, pp. 488 – 501.

Abstract: This paper presents a synchronic and diachronic computer corpus of Makarska littoral dialects. This corpus was created as part of the project to explore the ikavian neoštokavian dialects of the narrow coastal area in Croatian region of Dalmatia around the town of Makarska. The dialectological characteristics of the dialects studied are briefly presented first, followed by presentation of the digital system. The system is logically organized in first part as a corpus of literary texts created from 1729 to 1803 and digitally processed, and in the second part from the materials collected through dialectological questionnaires prepared and methodologically adapted as part of the creation of the Croatian Linguistic Atlas. Methods of collecting linguistic data, method of input into the digital form and methods and possibilities of data processing will be explained. Based on the input and search strategies within the system, the examples will prove the origin of the dialects of the Makarska littoral to be that of the ikavian neoštokavian dialect described in the dialectological literature. This computer-based principle of work is a novelty in Croatian dialectology which has not been digitally processed so far and offers a basis for future dialectological research. This platform can be used in order to shorten the time of data processing and to analyse them more systematically and more efficiently. So far, there has been no such digital repository for any Croatian speech. This project represents a thorough synchronic and diachronic study of one rounded language area.

Keywords: spoken corpus, corpus design, computer corpus, dialect corpus, dialectology, štokavian

1 INTRODUCTION

In 2016, a scientific research project was registered in Croatia with project manager Ivana Kurtović Budja, PhD. The project applied to the Croatian Science Foundation for its funding, and its aim was to conduct dialectological research at the designated points and to record material that will be archived physically and digitally. The default points are located in the Makarska littoral in Dalmatia, Croatia's southern region. The dialects of this area belong to the štokavian dialect, one of three Croatian

dialects (štokavian, čakavian, kajkavian). The characteristics of this dialect with confirmations obtained on the ground will be briefly described in the paper, and the entire material will be computer-processed and entered into the synchronic and diachronic computer corpus which will be accessible online and searchable in synchronic and diachronic mode and also searchable by different language criteria with audio recordings from native speakers.

2 RESEARCH OBJECTIVE AND METHODOLOGY

The aim of the project is a study from the historical linguistic aspect which will monitor specific Makarska littoral linguistic lines. The monitoring process is based on the corpus of old texts from the Makarska littoral, therefore, it will enable the analysis of continuity of the language from the oldest written monuments until nowadays. In addition, a sociolinguistic survey on the attitudes the young speakers have towards the local dialect and a dialect analysis according to different age groups and education levels will be conducted. The dialects of Brela, Podgora, Makarska, Igrane, Zaostrog, Živogošće, Baška Voda, Gradac, Promajna, Tučepi, Drvenik, Drašnice were studied. Apart from them, as control points, the dialects of Žrnovnica, Zadvarje, Rašćane, Maslinica, Sumartin, Sućuraj, Račišće, and the three Croatian idioms in Molise (Italy) on phonological, morphological, syntactic, and lexical level will be conducted. Informants, i.e., speakers of organic idioms, i.e., examinees, were audio recorded and questionnaires containing demographic data on informants were manually filled in. Each speech was then dialectologically described, materials were prepared for the synchronic part of the corpus, as were representative samples of the recorded natural idiom of every location to be used as a material basis for the spoken language corpus. Selected texts from the Makarska littoral, created from 1729 to 1803, were prepared for the text corpus.

The research was organised in several phases and at several levels; the dialectological research is carried out by finding reliable informants who are examined according to the standardized questionnaire for the Croatian Linguistic Atlas. All of the recording is approved by the informant with the statement of approval for participation in the project. Sociolinguistic research is conducted using also the method of field research, filling out questionnaires and processing materials. Historical linguistic and textological research is conducted on the basis of scanned and transcribed materials. Together, everything is prepared for entry into the synchronic and diachronic corpus of Makarska littoral dialects. The obtained results will be computer-processed and searchable according to phonological, morphological, and syntactic characteristics from a synchronic and diachronic aspect, which will demonstrate the continuity or discontinuity of each of the idioms.

3 AREA OF RESEARCH: GEOGRAPHICAL AND DIALECTOLOGICAL FEATURES

3.1 The Makarska littoral

In history, the Makarska littoral was a closed geographical and political unit because of its geographical location. It extends along a narrow coastal area of sixty kilometers in the Croatian region of Dalmatia.

3.2 Dialectological features

The Makarska littoral dialects belong to ikavian neoštokavian dialect [1] (Fig. 1). This dialect borders the eastern Herzegovinian and the neoštokavian ijekavian dialect on the south, and on the sea, just opposite to the Makarska littoral, there is the south-čakavian or ikavian čakavian dialect. The western Bosnian-Herzegovinian dialect, i.e., the ikavian štokavian dialect, is one of the seven štokavian dialects [2]. The largest ikavian štokavian unit consists of western Herzegovina, the Dalmatian mainland, a part of Lika and parts of western and central Bosnia. Today's borders of this dialect were set after the 15th century, in the third period of Croatian language development [3], when Turkish breakthroughs triggered large migrations in the population. For these nonlinguistic reasons, the dialects of today's ikavian štokavian dialect span larger and smaller unconnected regions. It was part of the western-štokavian dialect of the Croatian language, which fell apart after the Ottoman conquest.

The neoštokavian dialects are marked by the so-called neoštokavian language innovations in many dialects. It is about the neoštokavian accentuation (four-accent system), and mainly a systematic transfer of falling accents toward the beginning of the word and to proclitics (Nsg. *òko*, Lsg. *ù oku*). The letter *l* in final position gave way to *o* or *a* (inf. *biti*, r. pr. m. sg. *bȉja*), the letter *h* is lost (in standard Croatian *hrâst* ('oak'), in dialect Nsg. *râst*). Jat is ikavian (in standard Croatian *mlijéko* ('milk'), in dialect Nsg. *mlíko*). The researched dialects are šćakavian (in standard Croatian *iskati, tražiti* ('to look for'), in dialect 3. pl. praes. *ȉšĉū*). The letter -m in final position in suffixes and in inflectional words changed to –n. Numerous romanisms as well as turcisms are present in the vocabulary [4]. The group ra has two versions; *ra* (in standard Croatian *râsti* ('to grow'), in dialect inf. *râst* (Baška Voda)) and *re* (in standard Croatian *râsti* ('to grow'), in dialect inf. *rêst* (Brela)). The phoneme *h* is lost or rare mainly everywhere, and usually switches to *j* (in standard Croatian *grijêh* ('sin'), in dialect Nsg. *grîj*), v (in standard Croatian *bùha* ('flea'), in dialect Nsg. *búva*) or disappears in the initial position (in standard Croatian *hlâd* ('shade'), in dialect Nsg. *Lâd*) [5]. All these more distinctive features, as well as many other specific features for this dialect, can be attested for in the corpus, which achieves one of its objectives to make it digitally easily available for dialectological research.

Fig. 1. Map of štokavian dialect with the Makarska littoral marked as follows ([6], [7]) (dialects in the map legend from top to bottom: neoštokavian ikavian, neoštokavian ijekavian, neoštokavian ekavian, slavonian, eastbosnian, zetski)

4 DESIGNING THE SYNCHRONIC AND DIACHRONIC COMPUTER CORPUS OF MAKARSKA LITTORAL DIALECTS

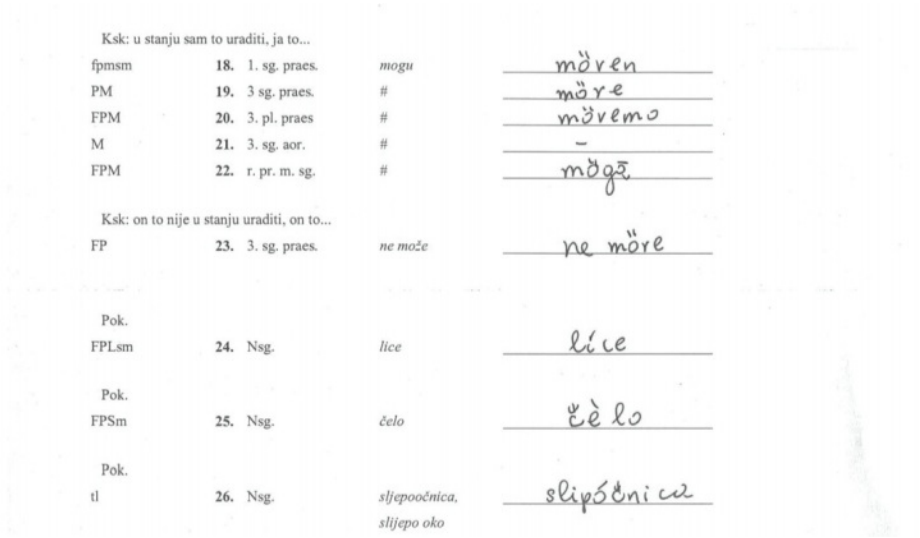
4.1 Collecting materials

The material constituting the corpus consists of data collected from the questionnaire for the Croatian Language Atlas, old texts from the area of the Makarska littoral, and the control texts of other dialects. The texts currently in the digital corpus are Hvarkinja by Martin Benetović (around 1550–1607), Bogoslovje diloredno by Antun Kadčić (1686–1745), Deset pokorni razgovora by Ivan Jozip

Pavlović Lučić (1755–1818) and Sarod Rakitichah by Petar Rakitić. The project contains further old texts which have been processed in optical character recognition program (OCR), then manually examined for errors. Those texts are Razgovor ugodni naroda slovinskoga (1756) and Korabljica (1760) by Andrija Kačić Miošić, Tridentinskoga sabora naredbe (1790), Commentarii morales (1793) and Dvi bogoljubne pofale (1803) by Ivan Jozip Pavlović Lučić. The basis for having those texts in corpus is the birthplaces of the authors which are in Makarsko primorje region. Exceptions are Hvarkinja by Martin Benetović and Sarod Rakitichah by Petar Rakitić. Those two authors are not from the region, neither are the texts written in the dialect of Makarsko primorje region. They have been chosen as control texts. Sarod Rakitichah is štokavian text of other area, not the same as in Makarsko primorje. Hvarkinja is a renaissance commedia ridicolosa but with many features common to commedia erudita. It has been chosen as a control text of another dialect but also because one of the characters in the drama is from Makarsko primorje region and speaks in its own vernacular. The idea for organizing the old text corpus, selection of texts and all the major work has been done by Jurica Budja, PhD, an associate on the project. Budja has also reedited Deset pokorni razgovora and wrote the introductory study. The plan is to do the same with all the old texts, i.e., books. All of the mentioned texts are part of the all-Croatian corpus, a much bigger set of texts from different authors, regions, dialects, and periods.

4.1.1 Language materials of the Croatian Language Atlas

Croatian local dialects were explored for multilingual atlases: the General Slavic Linguistic Atlas (28 points), the European Linguistic Atlas – Atlas linguarum Europae and the Central South Slavic Dialectological Atlas (called the Croatian-Serbian Dialectological Atlas) (236 points). In 1996, the Croatian Language Atlas (HJA) project was initiated at the Institute of Croatian Language and Linguistics, for which a network of 399 points of Croatian dialects (101 čakavian, 110 kajkavian, 188 štokavian) is envisaged [8]. The Croatian Language Atlas is a set of descriptions of Croatian dialects. The descriptive material was collected by systematic field work, that is, audio and written recording of language status at points in Croatia and locations outside Croatia in which Croatian language is spoken. Researchers, most often dialectologists, in conversation with the informants record linguistic characteristics which are agreed in advance and according to which the field conditions on the phonological (F or f), prosodic (P or p), morphological (M or m), word formation (T or t), lexical (L or l) and syntax (S or S) level can be closely monitored (Fig. 2).


Ksk: u stanju sam to uraditi, ja to...

fpmsm	**18.**	1. sg. praes.	*mogu*	mȍven
PM	**19.**	3 sg. praes.	#	mȍre
FPM	**20.**	3. pl. praes	#	mȍvemo
M	**21.**	3. sg. aor.	#	–
FPM	**22.**	r. pr. m. sg.	#	mȍgȕ

Ksk: on to nije u stanju uraditi, on to...

FP	**23.**	3. sg. praes.	*ne može*	ne mȍre
Pok. FPLsm	**24.**	Nsg.	*lice*	líce
Pok. FPSm	**25.**	Nsg.	*čelo*	čȅlo
Pok. tl	**26.**	Nsg.	*sljepoočnica, slijepo oko*	slipȍčnica

Fig. 2. The sample page of dialectological paper questionnaire (question asked the informant, linguistical level that is monitored, description of a word word, word told by informant)

The research, based on the Croatian Language Atlas method, is a basis also for the study of the Makarska littoral dialects. The situation is recorded with a questionnaire containing 2122 lexical data divided into 15 chapters covering different areas related to man, his environment and occupation:

- man and body parts
- family
- folk costumes (male and female)
- house and objects in the house
- diet and preparing meals
- grain farming and processing
- domestic animals
- birds and domestic poultry
- wildlife
- trees and fruits
- vegetables and flowers
- diligence
- land, water, natural phenomena
- social life
- supplement.

The supplement examines onomastic material, some adjective expressions that do not correspond to any of the previous chapters (e.g., sweet, thirsty, solid…), names for colours, pronouns, cardinal and ordinal numbers, adverbs (e.g., up, down, over, from, where, today…), verbs to be and to have in certain forms.

The words are also organized in headwords, that is, several different categories are required for a single flectional word (for flectional words forms in different cases, numbers, and for verbs different person, time, number and mood). A grammatical level is recorded for each entry, that is, showing why a particular word was even chosen to enter the questionnaire. Each headword is accompanied by a question that helps the examiner to obtain the required word (e.g., if the word head is requested, the examiner asks a question: what is the name for (with the abbreviation "Ksz"), and then points (with the abbreviation "Pok") to their head).

It is also possible that the examiner begins the sentence and does not finish it by asking the examinee to complete it logically. The missing word in the examiner's sentence is precisely the one to be entered in the questionnaire (e.g., It's dark in the room; nothing can be... and the examinee is expected to say the word seen in the form present in the idiom). Naturally, such way presents obstacles because it is not always easy to obtain the desired response, the long-term study creates fatigue in the examinee and examiner, and if the examiner is not the native speaker of the examinee's idiom, or if he speaks at the standard, the examinee will spontaneously start to approach them in their idiom. This can provide compromised data. That is why the strategy of an unstructured conversation is often used, i.e., a spontaneous speech of the examinee, who is asked to talk about their youth, activities, customs, to tell a story, anecdote… is recorded. Such way provides much more material than previously planned. Also, a material which serves as an intangible heritage is obtained because it contains valuable information about life in certain parts of the country.

4.1.2 Sociolinguistic and demographic data

The questionnaire contains detailed metadata that help classify the speech and describe the examinee and the situation in which the study was carried out, for example to which dialect the speech belongs, where the survey was conducted. In particular, the features of the site are recorded in detail:

- approximate number of inhabitants
- name of neighbouring places
- which city serves as a direct centre of the surroundings (economic, administrative, etc.)
- where children go to school, in which companies the adults work
- religion of the population
- what is the occupation of most inhabitants and whether it has always been the case

- inhabitant name for the place and neighbouring places
- spread of surnames over the place (including family nicknames)
- the most common male and female names
- names of streams, hills, rivers, springs…
- is there a proverb, sentence or gesture with which the inhabitants of neighbouring places mock each other
- what are the places around that have the same dialect
- what are the places around that have a different dialect
- origin of the population in the town
- to which animals do people give names
- enumerate plants
- phonological description of the place.

The following part records important details about the examinee, the personal ones and those by which it can be seen how their speech retained the features of the old expression or was changed under the influence of education, media, and migration:

- name and surname
- family nickname
- year of birth
- place of birth
- father and mother's place of birth
- number of ancestors born in the informant's place of birth
- nationality and religion
- the place which the spouse is from and how long they've been married
- where the informant lived in the past
- are they literate and to what extent
- data on education
- what is the degree of their intelligence
- what are their interests (past and present)
- how do they behave in relation to the conducted study (whether they show initiative or are they restrained, fearful, etc.)
- how much are they influenced by the standard Croatian language
- do they make special efforts to speak with the organic idiom (without the influence of the standard language) or are they trying to speak "lordly"
- what are their speech organs and articulation like.

4.2 Organisation of material in the computer corpus

The structure of the corpus consists of two separate parts, data collected from the questionnaire for the Croatian Language Atlas (HJA) and old texts (Fig. 3). The two are organised in two separate databases and are subjected to separate operations.

The HJA database is organized in such a way that the user (investigator) is connected to the examinee, and through the examinee the connection continues to the place where the speech is investigated. On the other hand, the user is related to the answers, i.e., any data entered in the questionnaire. Separate data is a part of the grammatical structure (division) concerning one word (all grammatical forms of flectional word). The division is related to one question, that is, a sentence by which the examiner receives information about the division, that is, about the initial grammatical category from which the entire grammatical structure of a headword is developed.

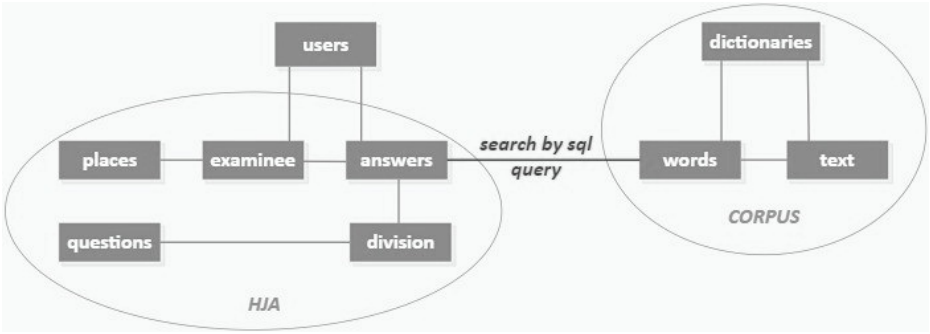


Fig. 3. The organisational scheme of computer corpus

The second database contains a corpus of old texts. This database contains complete texts and also all the words separated from them. Tokens were obtained by tokenization to the level of words, with punctuation marks neglected and not entered into the database as tokens.

Separate operations can be conducted on each database; content input and content analysis. However, to achieve a diachronic aspect of the corpus and enable diachronic analysis, it was necessary to enable communication between the two databases in order to compare historical and contemporary language data. Since these are separate databases, they are interconnected by sending SQL queries from one database to another and vice versa. The process of sending SQL queries for diachronic analysis is explained at the end of chapter 4.2.1 when comparing questionnaire data and old text corpus data.

4.2.1 The material from the questionnaire for the Croatian Language Atlas

The material collected through the questionnaire is transferred to the computer corpus on a website that has a data entry form with the same structure as the paper version of the questionnaire. Before entering the main part of the questionnaire – dialectological material or words – data on the place where the speech is examined and data on the informant are entered.

In the questionnaire, the phonetic characteristics of the phonemes are not recorded, but only the accent characteristics of vowels and syllabic r: quantity (long, short) and tone (falling, rising). Therefore, accentuation in ikavian neoštokavian idioms implies four accents with distribution as in the standard Croatian language, but with some exceptions [7]. This means that there are four different accents: short-falling –̏, short-rising – ̀, long-falling – ̂ and long-rising – ́. Some vowels that are not stressed but are long are marked with length – ̄. When entering each word, it is possible to obtain a drop-down box in which all special characters are offered to enter the correct accent or length (Fig. 4).

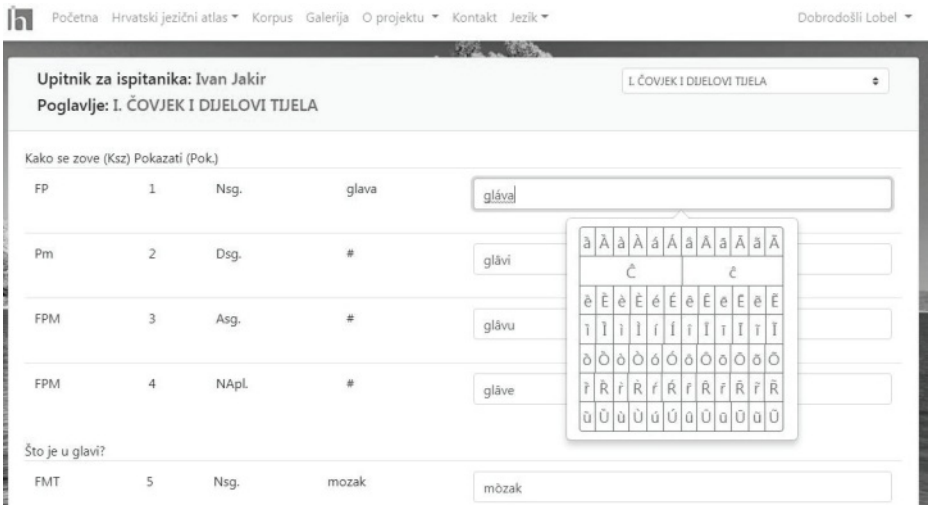


Fig. 4. Inserting the data with accent entry possibilities

The database was created in SQL and therefore data entry and review have a slight flaw. Some phonemes are specially written; č, ǯ, ǵ, Ĺ, ń. However, SQL does not support encoding special characters corresponding to their phonological transcription. Therefore, these signs are entered as č → ĉ, ǯ → dž, ǵ → đ, Ĺ → lj, ń → nj. They reflect the dialectological notation tradition from the transition of the 19th to the 20th century, although the versions on the right side of the arrow are increasingly used today [9].

Entering data from the questionnaire provides a table in which each entity (the entire table line) is described by four attributes or features (table columns). The attributes of the table are the grammatical level for which the requested word was examined, the grammatical determination of the requested word, the word in its standard Croatian language form and number of how many dialectological results were entered for each standard language word (Fig. 5).

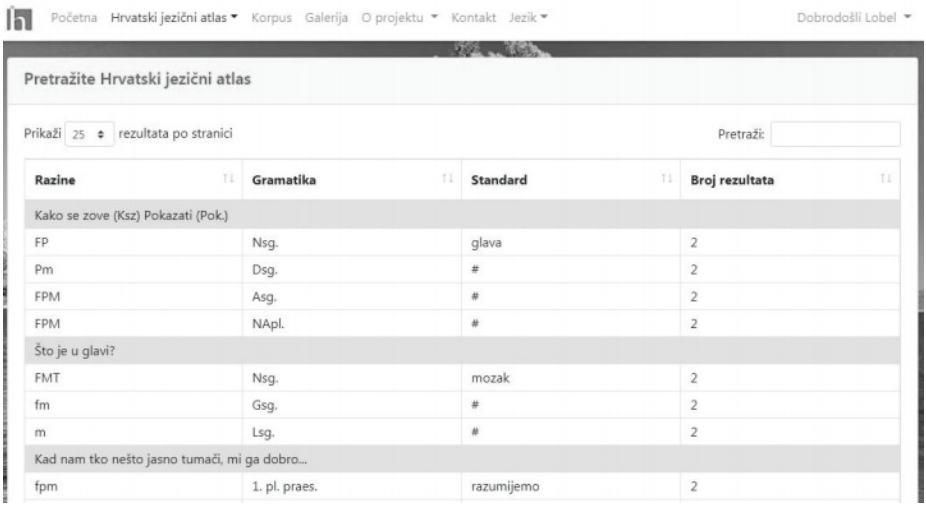


Fig. 5. Digitized dialectological questionnaire (columns: linguistical levels, grammar, standard form, number of results)

Clicking on an entity opens an overview showing separately all occurrences in all surveyed local dialects. The review is organised by providing metadata, i.e., a description of the entity, in the upper part. The chapter in which the word is included shall be indicated, the question assisting the examiner to obtain an answer, the grammatical level, the grammatical description and the standard version of the word. Below that there is a table that shows in which places the word appears, in which form and what its frequency is (Fig. 6). In addition to these data, one can find additional options in the last column of the table. By clicking on the note icon one can get an audio clip and hear how the native speaker pronounces the word, and by clicking on the book icon there is an insight into which old texts contain that exact word and what its frequency is. An overview of the content from the questionnaire is a synchronic overview of the local dialects, and a comparison with the old texts (book icon) is a diachronic overview (Fig. 7).

4.2.2 Old texts in corpus

The second part of the computer corpus consists of texts created from 1729 to 1803. They were selected as representative examples of the language of the surveyed area. The first published editions where put into the optical character recognition system (OCR). They were then tokenized – the text is divided into occurrences (Fig. 8).

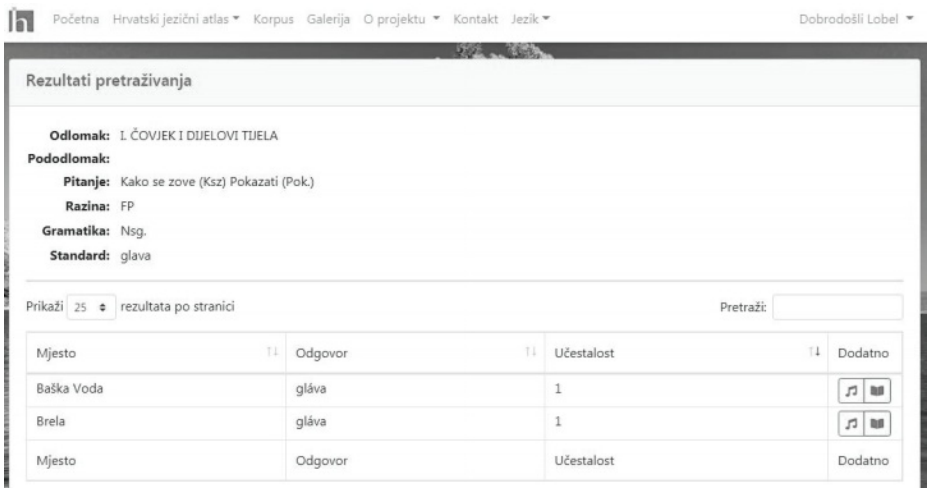


Fig. 6. Description of each entity and its variants in local dialects (columns: place, answer, frequency)

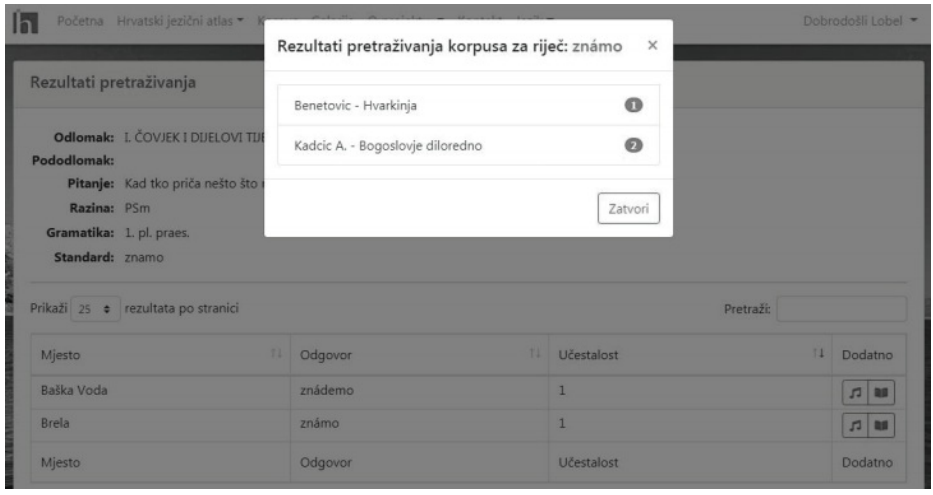


Fig. 7. Finding the word in old literary texts (for each word there is a number of occurrences in each of them)

That process enabled comparison of the text database with the questionnaire database at the level of words. This gives a diachronic insight into the similarities and differences between the historical and present language forms, that is, the development of words at grammatical levels can be historically monitored. Each text in the corpus can be viewed in its entirety, in the form of text rather than tokenized units. In this case, the text is divided into pages, as it is divided into pages in the original.

Br.	Riječ	Broj pojavaka	Br.	Riječ	Broj pojavaka
1	k	3	2	ka[1]	3
3	kacaj	1	4	kad	3
5	kada	3	6	kadcich	1
7	kadgod	1	8	kadgodir	1
9	kadijom	2	10	kadimo	1
11	kadčića	1	12	kaifāsu	1
13	kajat	2	14	kajatī	1
15	kajaše	1	16	kaje	1
17	kajem	2	18	kaješ	1
19	kajo	1	20	kaju	2

Fig. 8. The occurrences in old text corpora (columns: number, word, number of occurrences)

It is also possible to perform logical operations over different sets of texts. The system offers to form two sets of texts. Following the formation of the two sets, possible operations are provided: view set A, view set B, A–B, B–A, UNION(A,B), INTERSECTION(A,B), XOR(A,B).

5 CONCLUSION

The synchronic and diachronic computer corpus of Makarska littoral dialects is a corpus accessible online, divided into two larger data groups. One group consists of data collected from the dialectological questionnaires for the Croatian Linguistic Atlas – a list of points where the dialectological situation is examined. Based on it, the phonological descriptions of Croatian dialects are made, which constitutes one synchronic level. These descriptions are used to create dialectological maps and to monitor the distribution status of certain language phenomena. The second group of data is a corpus of linguistically representative old texts from the Makarska area, which serve as a basis for comparison with the current language situation. The texts present new synchronic level of their time each, but together they make a diachronic cross-section of literary texts. Finally, old texts and materials from the questionnaires together can be compared by applying principles of data research and comparison to investigate synchronic and diachronic differences and similarities. This computer system is the first example of the computer processing of dialectological questionnaires in Croatia. Although it was drafted within the specific project which

explores the dialects of a smaller area and a single dialect, the principles of computer processing and subsequent research are applicable for all Croatian dialects and the processing of all questionnaires filled in so far and for all questionnaires to be done. Indeed, in the process of collecting materials it is now possible to skip the step of manual paper questionnaire filling out but the materials from the field can be immediately entered into the computer system which can later generate paper-based edited questionnaires. It is also possible to extend the text corpus, which may cover different periods and thus provide insight into the language phenomena of certain times and places. Further upgrading of the system will enable morphosyntactic processing of words.

References

[1] Štokavski dijalekat. In Enciklopedija Jugoslavije. (1960). Leksikografski zavod FNRJ. Zagreb.

[2] Lisac, J. (1998). Štokavski i torlački idiomi Hrvata. In Najwse dzieje języków słowiańskich. Editor Lončarić, M. Opole (Poland). Uniwersytet Opolski – Instytut Filologii Polskiej.

[3] Brozović, D. (1970). O Makarskom primorju kao jednom od središta jezično-historijske i dijalekatske konvergencije. In Makarski zbornik 1, pages 381–405.

[4] Lisac, J. (2003). Hrvatska dijalektologija 1. Hrvatski dijalekti i govori štokavskog narječja i hrvatski govori torlačkog narječja, pages 50–71, Zagreb, Hrvatska.

[5] Govori Makarskoga primorja – sinkronija i dijakronija. Accessible at: http://elte2.fsb.hr:8000/gomapri, (Cited 20. 3. 2021).

[6] Edited by map in Hrvatska enciklopedija. (2002). Vol. 4, Fr–Ht, page 676. Zagreb. Leksikografski zavod Miroslav Krleža; graphic design by Tomislav Kaniški.

[7] Croatica. Povijest hrvatskoga jezika. (2018). Vol. 5, page 540. Zagreb.

[8] Kurtović Budja, I. (2020). Upitnici za Hrvatski jezični atlas i Zlatna formula hrvatskoga jezika ča-kaj-što. In Conference abstracts of Computational Linguistics and Golden Formula of Croatian Language. Osijek.

[9] Brozović, D. (1981). O fonetskoj transkripciji. In Fonološki opisi srpskohrvatskih/hrvatskosrpskih, slovenačkih i makedonskih govora obuhvaćenih opšteslavenskim lingvističkim atlasom, pages 17–25.

sciendo
DOI 10.2478/jazcas-2021-0046

MAPKA: A MAP APPLICATION FOR WORKING WITH CORPORA OF SPOKEN CZECH

HANA GOLÁŇOVÁ – MARTINA WACLAWIČOVÁ
Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic

GOLÁŇOVÁ, Hana – WACLAWIČOVÁ, Martina: Mapka: A map application for working with corpora of spoken Czech. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 502 – 509.

Abstract: A new interactive map-based web application named Mapka was published by the Institute of the Czech National Corpus in 2020. It aims to serve linguists, as well as schools and the general public, and it features various functions described in this paper. Mapka was designed as a supplement to the CNC spoken corpora, starting with the DIALEKT corpus (more to come in the future). Its main function is to display various types of territorial division (primarily in terms of dialect, but also administrative) and networks of localities associated with the corpus. The main dialect regions are provided with overviews of their typical dialectal features and two samples of dialectal discourse – one slightly historical and one contemporary. The application offers the possibility of searching for municipalities, plotting the points on the map and creating a custom map. The paper concludes with future prospects concerning an enhanced and improved version of the application.

Keywords: corpus, map, Czech language, spoken language, dialect

1 INTRODUCTION

In July 2020, the Institute of the Czech National Corpus published a new tool: a web application named Mapka [1]. It is an interactive map-based application primarily designed as a supplement to the spoken corpora of the Czech National Corpus, however it features various functions beyond this framework. The Mapka application is intended to serve both linguists and the general public. It is accessible without registration and it is available at: https://korpus.cz/mapka/.

Currently, its main function is to display the various types of language boundaries (and additionaly administrative borders) and nets of localities represented in the DIALEKT corpus ([2], [3], [4]) using an interactive map of the Czech Republic. The next phase, planned for the current year of 2021, will include (amongst other things) additional data from other spoken CNC corpora, e.g., ORTOFON [5] and ORAL [6]. The application includes presentations of characteristic features of the main Czech dialect regions illustrated by authentic speakers' utterances – slightly historical, as well as contemporary ones. Users are enabled to search for municipalities, add these points to the map, and create their own map.

The goal of this paper is to showcase the current version of the Mapka application, introduce its possible uses, and outline future prospects.

2 FEATURES AND FUNCTIONS OF THE MAPKA APPLICATION

2.1 Territorial division

The Mapka application displays various types of territorial division on the background map. The most important of these is the dialect-based territorial division of the Czech-speaking language territory, i.e., 10 regions (Central Bohemia, Northeast Bohemia, West Bohemia, South Bohemia, Bohemian-Moravian transient region, Central Moravia, East Moravia, Silesia, Bohemian borderland, Moravian and Silesian boderland) including the Bohemian borderland and the Moravian and Silesian borderland, although they do not belong to the group of traditional dialect regions. In this context, borderland refers to the historical area defined by the former numerical prevalence of the German-speaking population, massive population relocation after World War II, and the lack of a traditional Czech dialect substrate. Moreover, the Czech-speaking language territory is not only the territory of the Czech Republic, but a few localities – Czech language islands belonging to the dialectal region of Northeast Bohemia or Silesia – are located in Poland.

If needed, it is possible to choose a mode showing even more detailed dialect categories: dialect region / *nářeční oblast*, dialect subgroup / *nářeční podskupina*, dialect area / *nářeční úsek*, dialect type / *nářeční typ*. This detailed division is important mainly for the area of Moravian and Silesian dialects. Concerning the Bohemian territory, these detailed dialect categories can be found almost solely in the border areas of its dialectal regions. The Mapka application shows the exact position in the dialect system and a numeric code (used by Bělič [7]) for each dialect area of any type. The system of territorial dialect division employed in both the DIALEKT corpus and Mapka is based on Bělič's approach, *the Czech Linguistic Atlas* ([8], [9], [10]), *the Encyclopaedic Dictionary of Czech* [11] and its presentday online version: *The New Encyclopaedic Dictionary of Czech* [12].

Besides this, the map also provides an option of displaying the boundaries of the Czech Republic's administrative units, i.e., districts / *okresy* or regions / *kraje*.

The basic background map can be easily switched from to the relief map which, e.g., enables comparing natural and dialectal boundaries and discovering their connections.

The process of developing the application required a thorough revision of dialect regions delimitation, mainly particularization of their borders. It was based on a large amount of dialect monographs, studies, and consultations with dialectologists. Additional verification was made using the statistical lexicons of municipalities ([13], [14]). As far as the boundary delimination is concerned, problems of transition zones must be mentioned. In dialectology, the transition

zones are defined by a gradual decrease in the number of dialectal features typical for dialectal subgroup, area, or type. Classification of some municipalities can be difficult, for there can be a lack of certain features typical for one dialect area on the one hand, and on the other hand, features typical for the neighbouring area may be missing. For example, the municipality Brodek u Přerova belongs to the Core Central Moravian Subgroup (*hanácká*), however the phonetic changes y > e and u > o typical for the subgroup do not occur in the above mentioned municipality. In spite of this, it cannot be subsumed into another subgroup, as it does not evince appropriate dialectal features. In the future version of the application, this problem will probably be solved by graphic highlighting of the transition zones along the borders.

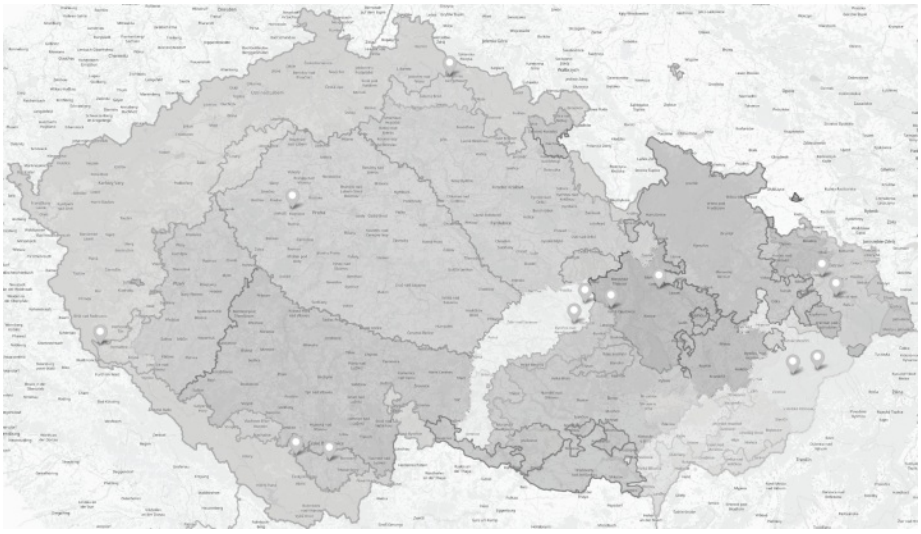


Fig. 1. Detailed dialect division of the Czech-speaking language territory. The samples of dialectal discourses are plotted as white pins

Currently, new types of territorial division based on historical data are being prepared to be added to the Mapka application. One of them is plotting the historical Bohemian-Moravian border and the Moravian-Silesian border, as they were formed at the end of the 12th century and stabilized in the 14th century. These borders have not been used by administrative authorities since 1948, nevertheless, they have a huge historical significance. Another new layer of the map will display the German language islands in the Czech Republic. These are eight areas where there was a historical numerical prevalence of the German-speaking population and the German language was used. They are located mainly in Moravia, some of them have

urban characteristics, some of them rural. They disappeared in the 20th century – partly in the first half of the century, partly after the World War II.

2.2 Overviews of dialectal features and samples of dialectal discourses

The Mapka application encompasses several overviews of typical dialectal features pertaining to the main three regions of the Czech Republic (Bohemia, Moravia and Silesia) and eight dialect regions of the Czech-speaking language territory (see above). The overviews are mainly focused on phonological and morphological features as they are fundamental for the dialect division. Examples of dialectal phenomena were primarily selected from the current version of the DIALEKT corpus, but some examples were supplementarily taken from transcripts which will be published in the upcoming version of the corpus.

Furthermore, each of the eight main dialectal regions is illustrated by two samples featuring authentic dialectal discourses chosen from the DIALEKT corpus. The samples consist of an audio recording and its two transcripts – dialectological (based on the Rules for the Scientific Transcription of Dialectological Records of Czech and Slovak [15]) and orthographic. The samples were chosen in order to demonstrate the most typical dialectal features of the given regions. The recordings of the DIALEKT corpus are divided into two time strata, hence for each dialect region, one sample was selected from the older stratum of dialect material (the period between the late 1950s and the 1980s) and one from the newer one (from the 1990s up to the present day). The samples were chosen so that both recordings representing a particular region were made in the same municipality or a neighbouring location. This guarantees maximum comparability of discourses and users can trace changes of the local dialect in time. Each sample is followed by an analysis that describes relevant dialect features (phonological, morphological, syntactic and lexical) occurring in the discourse. In the future, samples for other prominent dialectal sections and types will be added. Our goal is to capture the variability of the traditional regional dialects as much as possible.

2.3 Mapka as a supplement to the DIALEKT corpus

While designing the Mapka application, the primary aim was to create a supplement for the spoken corpora of CNC, that would integrate data from these corpora with a map-based interface. For the time being, Mapka serves as a supplement for the DIALEKT corpus. Above all, municipality networks that have certain connections to the DIALEKT corpus can be displayed on the background map. For example, users can observe a network of municipalities where recordings included in the current version of the DIALEKT corpus were produced. It is possible to display all of the concerned municipalities or to choose either the network of localities, where recordings from the older time stratum were produced, or the network of localities bound with the newer stratum. Another option is to visualize

the network of all municipalities where overall data collection for this corpus (published recordings as well as unpublished so far) took place. A special bonus is the possibility to display the network of research localities of the Atlas of the Czech Language. All these networks can be visualized simultaneously and compared. Users can choose which one of them will be displayed above all the others. If needed, the work with the map can be restricted to a particular dialect region or regions and the others will not be considered.

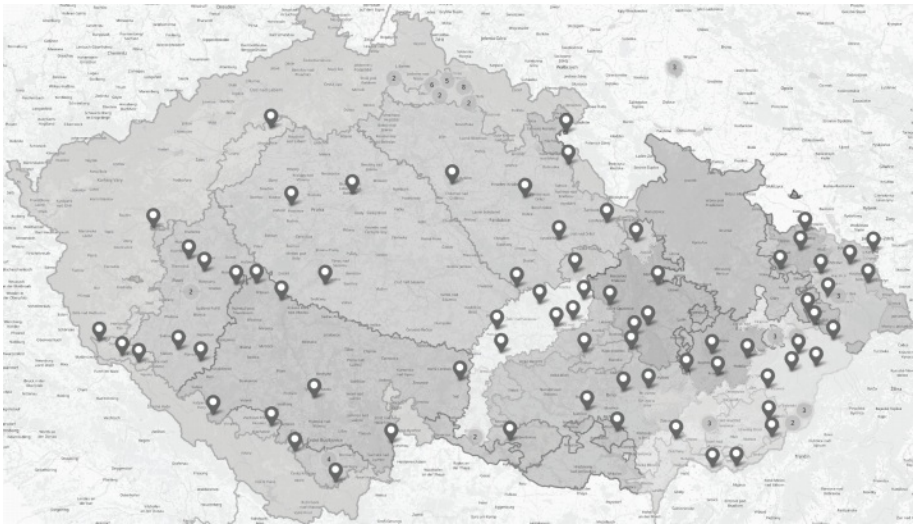


Fig. 2. Network of municipalities where overall data collection for the DIALEKT corpus took place. The dark grey pins containing numbers mark areas with a larger amount of points

2.4 Searching and creating your own map

The Mapka application offers the option to search for municipalities / parts of municipalities in the Czech Republic. The cadastral boundaries of the looked-up municipalities are visualized on the map, in order to be clear which parts belong to a certain municipality. Users can display information about the position of the municipality in the system for the division of dialectal regions.

The application enables users to proceed to plot these points on the map and create their own map. They can choose colours from the colour spectrum for differentiation of various groups of points.

The resulting maps can be downloaded and printed. In the future version, users will also be able to save their map with plotted points and continue working on it later, after loading the application again. We hope it will prove to be useful for linguists doing research or preparing their own map to illustrate their monographs or studies.

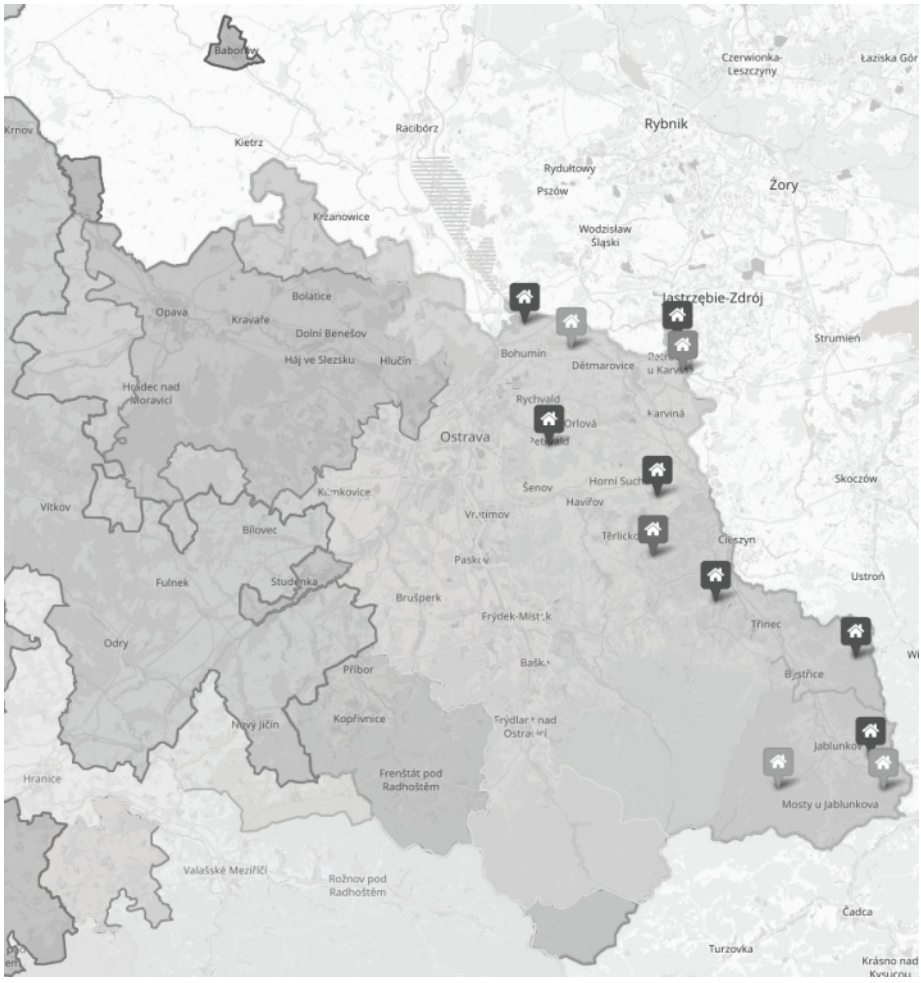


Fig. 3. Example of creating a map. Markers of four different colours identify four groups of locations in Silesian-Polish dialect subgroup. Each colour is bound with a certain dialect monograph and markers refer to the locations where language samples originated from

3 FUTURE PROSPECTS AND GOALS OF THE MAPKA APPLICATION

An enhanced version of the Mapka application is being prepared and hopefully will be published soon. Some of the planned innovations have been mentioned above, but in this section, we would like to sum up all future prospects related to Mapka.

As far as the DIALEKT corpus is concerned, new samples featuring authentic dialectal discourses for other prominent dialectal sections and types will be added to the application.

Considering the Mapka application was designed as a supplement to all of the spoken corpora of CNC, data from spoken corpora such as ORAL or ORTOFON will be included in the application prospectively. The data will encompass e.g. word counts, recording counts, statistics about the speakers, networks of locations of the speakers' childhood residence (until 15 years of age) or places of their longest residence, or networks of locations where recordings were made. Collections of data will probably be different for each corpus, since the corpora include different types of sociolinguistic metadata. We are planning on incorporating samples of authentic discourses chosen from the ORAL or ORTOFON corpora. The samples could be manually chosen and prepared or could be automatically generated random samples or both.

When speaking of creating a personalised map, the possibility to save the user's data will be integrated into the application.

We are considering creating an entertaining linguistic quiz focused on dialects or spoken language in general and its incorporation into the application. It could be attractive for the general public or schools.

The primary goal of the Mapka application is to capture the variability of spoken language across the Czech-speaking language territory. Taking the innovations planned for the next version of the application into account, we will be able to follow many particular aims such as to show differences between dialects captured by the DIALECT corpus and everyday spoken Czech language captured by other spoken corpora (e.g. ORAL and ORTOFON), differences between the language of the oldest generation of speakers and the other generations, between urban and rural speech or between monological and dialogical discourse. The application will help during the research of various aspects of spoken language, i.e. phonology, morphology, syntax, lexis, pragmatics or dialogue construction. The application is expected to serve language experts, all levels of schools as well as amateurs from the general public.

ACKNOWLEDGEMENTS

This paper resulted from the implementation of the Czech National Corpus project (LM2018137) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

References

[1] Goláňová, H., Waclawičová, M., and Pejcha, J. (2020). Mapka: Mapová aplikace pro korpusy mluvené češtiny. Version 1.0. Praha: ÚČNK FF UK. Accessible at: http://korpus.cz/mapka.

[2] Goláňová, H., Waclawičová, M., Komrsková, Z., Lukeš, D., Kopřivová, M., and Poukarová, P. (2017). DIALEKT: nářeční korpus, verze 1 z 2. 6. 2017. Praha: ÚČNK FF UK. Accessible at: http://www.korpus.cz.
[3] Goláňová, H. (2015): A new dialect corpus: DIALEKT. In K. Gajdošová and A. Žáková (eds.), Proceedings of the Eight International Conference Slovko 2015 (Natural Language Processing, Corpus Linguistics, Lexicography), pages 36–44. Lüdenscheid: RAM-Verlag.
[4] Goláňová, H., and Waclawičová, M. (2019). The DIALEKT corpus and its possibilities. Jazykovedný časopis, 70(2), pages 336–344.
[5] Kopřivová, M., Komrsková, Z., Lukeš, D., Poukarová, P., and Škarpová, M. (2017). ORTOFON: Korpus neformální mluvené češtiny s víceúrovňovým přepisem. Praha: ÚČNK FF UK. Accessible at: http://www.korpus.cz.
[6] Kopřivová, M., Lukeš, D., Komrsková, Z., Poukarová, P., Waclawičová, M., Benešová, L., and Křen, M. (2017). ORAL: korpus neformální mluvené češtiny, verze 1 z 2. 6. 2017. Praha: ÚČNK FF UK. Accessible at: http://www.korpus.cz.
[7] Bělič, J. (1972). Nástin české dialektologie. Praha: Státní pedagogické nakladatelství, 463 p.
[8] Balhar, J., Jančák, P. et al. (1992, 1997). Český jazykový atlas 1, 2. Praha: Academia, 427, 507 p.
[9] Balhar, J. et al. (1999, 2002, 2005). Český jazykový atlas 3, 4, 5. Praha: Academia, 577, 626, 680 p.
[10] Balhar, J. et al. (2011): Český jazykový atlas Dodatky. Praha: Academia, 579 p.
[11] P. Karlík, M. Nekula, and J. Pleskalová (eds.). (2002). Encyklopedický slovník češtiny. Praha: Nakladatelství Lidové noviny, 604 p.
[12] P. Karlík, M. Nekula, and J. Pleskalová (eds.). (2016). Nový encyklopedický slovník češtiny. Accessible at: ttps://www.czechency.org/.
[13] Statistický lexikon obcí v zemi České. (1934). Úřední seznam míst podle zákona ze dne 14. dubna 1920, čís. 266 Sb. zák. a nař. Statistický lexikon obcí v republice Československé. Praha: Orbis, 643 p.
[14] Statistický lexikon obcí v zemi Moravskoslezské (1935). Úřední seznam míst podle zákona ze dne 14. dubna 1920, čís. 266 Sb. zák a nař. Statistický lexikon obcí v republice Československé. Praha: Orbis, 236 p.
[15] Dialektologická komise České akademie věd a umění. (1951). Pravidla pro vědecký přepis dialektických zápisů českých a slovenských. Praha: Česká akademie věd a umění, 5 p.

sciendo
DOI 10.2478/jazcas-2021-0047

L2 CZECH ANNOTATION FOR AUTOMATIC FEEDBACK ON PRONUNCIATION

RICHARD HOLAJ[1] – PETR POŘÍZKA[2]
[1] Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic
[2] Department of Czech Studies, Faculty of Arts, Palacký University, Olomouc, Czech Republic

HOLAJ, Richard – POŘÍZKA, Petr: L2 Czech annotation for automatic feedback on pronunciation. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 510 – 519.

Abstract: In this paper, we would like to provide a brief overview of the current state of pronunciation teaching in e-learning and demonstrate a new approach to building tools for automatic feedback concerning correct pronunciation based on the most frequent or typical errors in speech production made by non-native speakers. We will illustrate this in the process of designing annotation for a sound recognition tool to provide feedback on pronunciation. At the end of the paper, we will also present how we have tried to apply this annotation to the tool, what caveats we have found and what our plans are.

Keywords: pronunciation, L2, Czech, machine learning, neural networks, e-learning, annotation, speech recognition, automatic feedback, phonetics

1 INTRODUCTION

Over the last few decades, online language learning popularity has been growing rapidly [1]. There are dozens of e-learning applications for different languages. These include several tools focused on various languages, most notably Duolingo [2], and a large number of applications focusing on just one language or aspect of language, such as Ten Ta To [3] or CzechME [4].[1] The increasing worldwide popularity and importance of e-learning education have been accelerated even more by the current epidemiological situation caused by Covid-19 [5]. Despite this increasing importance, there is an aspect of language that does not receive as much attention in e-learning, this being pronunciation. This problem is even more critical in a less common L2 such as Czech.

2 LANGUAGE PRONUNCIATION FEEDBACK IN E-LEARNING SYSTEMS

Putting aside a few exceptions, basically the only way e-learning applications approach the teaching of pronunciation is by providing the possibility to play

[1] CzechMe was created as a result of TAČR TL01000342 – an adaptable mobile application for teaching Czech to foreigners. Both authors of the paper were participants in this project.

recordings of words and phrases. Some of those applications also provide the possibility to record users and play and compare their recording with an original record. In general, there is a lack of any feedback or lessons that would teach users how to attain correct pronunciation, or at least a certain pronunciation level. For the Czech language, we are aware of just two exceptions: CzechME and Duolingo. In the first case, there are several lessons focused on sound discrimination (differentiation) and pronunciation, however, the current version of CzechME does not provide feedback on the user's pronunciation. In the second case, an automatic speech recognition (ASR) system is used to transcribe a recording to text, which is then compared with a text that should have been pronounced.

Although some applications often use existing ASR systems (such as Google Cloud Speech or CMUSphinx) to convert speech to text in order to provide feedback to students, there is one big caveat when using ASR technology for learning pronunciation. ASR technology is designed to understand: even when the pronunciation is incorrect, it uses a language model [6] to guess what has been meant. This is a problem, since we receive the feedback that our pronunciation is correct even when it could have been more than just slightly wrong.

These types of tools are used across different L2 and it is apparently a state of the art solution for L2 pronunciation learning with one exception: a mobile application called ELSA Speak [7]. This pronunciation-only application provides an exhaustive amount of pronunciation exercises for English. It also includes a custom proprietary solution for evaluation of correct pronunciation and includes feedback to the user. The feedback is in the form of a speech sound which should be pronounced and the speech sound that the user actually pronounced. As far as we know, this is currently the technologically most advanced e-learning system for teaching L2 pronunciation, although there are still a number of issues. The system is only limited to segmental aspects, (level) of pronunciation and according to [8] the system still "often mistakenly identifies incorrect sounds as correct", thus the problem from ASR technology still partially remains. Another issue is with the feedback, which is limited to the speech sound inventory of English, despite the fact that the sounds pronounced by students often do not correspond to any sound in the target language (in this case English). The last issue leads us to the idea that we need more than a sound inventory of target L2 to create a successful system for providing feedback on the pronunciation of L2 (in our case Czech).

3 NON-NATIVE SPEECH RECOGNITION AND THE FEEDBACK APPROACH

The general idea of our approach is to include non-native sounds into an inventory of the speech recognition system, so we are not limited to the most similar speech sound from the language and thus we can obtain less distorted results of

actual pronunciation. Based on this recognition, we want to provide feedback to students that will tell what was wrong in their pronunciation and how to fix it. The feedback should not be in the form of the pronounced vs correct sound, which can be confusing for a student who usually does not know IPA. The form of feedback should be more explicit, for example, instructing students that their lips should be rounded or mouth more open, etc.

To achieve this target, we first have to collect a large amount of data and create an annotation system that will allow us to tell the differences between the speech sound that should have been pronounced and the speech sound that actually was pronounced. We will then need a tool based on annotated data and capable of recognising a speech sound and its corresponding annotation from recordings. In the first phase, we decided to test this approach on the individual speech sounds of L2 Czech.

3.1 Data: Collection and methodology

For the data collection, we had to take into account the technical aspects of the recordings, which were intentionally taken at varying levels of quality: (1) studio standard (44.1/16, wav); (2) compressed formats (mobile phone). Mobile recordings were used for the annotation and subsequent training to more closely match the quality of the recordings of the future mobile application.

187 foreigners – native speakers of 36 different languages – across all levels of language teaching (using the CEFR scale from A0 to C1) have been recorded thus far.[2] The speakers were recorded during Czech language courses for foreigners at the Summer School of Slavonic Studies at Palacký University Olomouc, as well as at the Center for Foreigners in Brno. All age categories from 18 to 73 years are represented, with the largest group being speakers under 40. The cumulative frequency is as follows: under 25 (66), under 35 (100), under 40 (136), under 50 (176), 50+ (187). In terms of gender, women predominate (114) over men (67) and over unknown (6).

The sample dataset contained isolated speech sounds, as well as two- to four-syllable words or phrases in which a given speech sound appeared in different positions (initial, middle, final) and in different phonemic contexts (vowels, obstruents, sonorants). The data was read twice by the non-native speakers – first with an instructor and then without any assistance. The students were asked to read all the Czech speech sounds in isolation at the end. Only part of the data (from 32 speakers, see below) – a set of segments with isolated speech sounds – has been used thus far to annotate the pilot testing of the recognition model.

[2] The teaching level with the dominance of the lexico-grammatical level does not have to correspond to the level of pronunciation. Representation was the following: A0–A1 (76 speakers), A0–A2 (112 speakers), others.

The design of the annotation system was based on a number of hypotheses and reflected (i) the phonetic basis of Czech, (ii) the phonetic specificities of foreign languages and the relevant phenomena and (iii) the most frequent pronunciation errors among foreigners learning Czech. These hypotheses were postulated from many years of experience with teaching foreigners by one of the authors. Deficiencies in the pronunciation of foreigners can generally be divided into a few different categories [9]:

(1) pronunciation of speech sounds that are not part of the Czech phonetic system, although the student is capable of pronouncing the Czech speech sound; these cases often stem from the written form of the language;

(2) pronouncing the speech sound is only problematic in certain positions or in close proximity to certain other speech sounds;

(3) the student is unable to distinguish two sounds – for speakers of Arabic, this can be [b] and [p], etc.;

(4) the speech sounds are not pronounced in a Czech style, such as when English-speaking students pronounce [p, t, k] with aspiration, etc.;

(5) the speech sounds cannot be pronounced by the student at all, not even approximately.

In creating the annotation system, we tried to take into account the "type and severity" of error, in the sense of: (1) slight deviations without compromising intelligibility – (2) deviations partially compromising intelligibility – (3) significant deviations compromising intelligibility (confusion of meaning, etc.). This categorisation could also be used as a way of providing the students with feedback.

3.2 Phonetic features that most often cause problems for foreigners

Certain speech sounds cause problems for foreigners regardless of their native language – they are difficult for practically everybody. At the segmental level, these are mainly the following phenomena (this is only a very brief and simplified list of the most common pronunciation issues):

- vowels – quantity: while it may be due to a lack of knowledge, certain foreigners may be applying what they are used to from their native languages; there is also the nasal production of vowels or diphthongs, "hard" pronunciation of [ɪ] or [u];

- consonants – the most difficult consonant for foreigners is the trill *ř* (whether in its voiced or unvoiced form: [r̝] [r̝̊]) and the laryngeal *h* [ɦ], which, although it exists in many languages, is not present in them in a voiced form;

- nearly all foreigners struggle with the consonants *ď* [c], *ť* [ɟ], *ň* [ɲ] (in contrast, speakers of Russian, Ukrainian or Azerbaijani frequently incorrectly soften the denti-alveolar *t*, *d*, *n*);
- pronouncing the syllable-forming consonants *l* [l̩] and *r* [r̩] – they are new to most foreigners and difficult to correctly articulate;
- issues with pronouncing Czech sibilants (*s* [s], *z* [z], *c* [t͡s], *š* [ʃ], *ž* [ʒ], *č* [t͡ʃ]) are also common;
- there are issues with distinguishing the voicedness of consonant pairs; tendencies to aspire in the pronunciation of plosives [p, t, k], articulatory issues with the lateral fricative [l].

Contextual or combinatorial phonetic phenomena are very important. The assimilation of voicedness in Czech can be, for example, a phenomenon new to many foreigners and pose issues for some; for some students, pronunciation difficulties are the result of a different articulation base or different assimilation processes (e.g., the tendency to use progressive assimilation, etc.).

4 ATTRIBUTIVE ANNOTATION SYSTEM

We created a formalised ATTRIBUTE–VALUE annotation system based on systematically categorised pronunciation errors from individual languages or language groups. The annotation label is divided by a colon into two main parts: (1) the part before the colon lists the speech sound that was supposed to be pronounced; (2) the attributes after the colon list (using the possible values of the given attribute) the deviations in pronunciation from the standard and the correct phonetic form of the speech sound. If the pronunciation is correct, only the part before the colon is used. In the case of incorrect pronunciation, any number of attributes can follow the colon (see below).

The annotation system specifies two groups of attributes: (1) *fixed*, which have a binary value of 0 or 1 for the phonological characteristics (quantity, voicedness) and (2) *variable*, with the possibility to add other values as needed (phonetic features such as palatalisation, etc.). There is a separate label for replacing one speech sound with another which has the format of X::Y where X = the desired speech sound and Y = the actually pronounced speech sound.

The *;err* tag denotes an unspecified pronunciation error. It can also be optionally supplemented with information on the acceptability of the non-normative pronunciation using the letters A or N to form *;errA* (acceptable) or *;errN* (not acceptable). The tagset labels for attribute values are unique and non-doubled, so ambiguity is not an issue.

Listed below are several examples, the format is always one speech sound per line (the meaning of the tag is explained in square brackets):

o:k1vN [short vowel *o* pronounced as long and nasalised]

a:vNvT [vowel *a* is nasalised with a hard pronunciation]

e [vowel *ɛ* is correct]

ou:vNkD_1 [both parts of the *o͡u* diphthong are nasalised, the first part is lengthened]

t':vR_t'j [the consonant *c* is pronounced in a segmented way with the inserted speech sound *j*]

Explanatory notes[3]:
attributes
k = quantity
v = non-normative pronunciation variants

*values of the **k** attribute*
K shortening
D lengthening

*values of the **v** attribute*
N nasalisation
R "segmented" pronunciation [supplemented by *aspects of value*]
T hard pronunciation

*aspects of values (can be assigned to any value of the **k** or **v** attribute)*
_1 error related to the first part of the diphthong
_2 error related to the second part of the diphthong
_xy *xy* represents the specific speech sounds in the segmented pronunciation

5 TESTING THE NON-NATIVE INDIVIDUAL SPEECH SOUND RECOGNITION

To test our annotation system, we decided to build a minimalistic tool for individual speech sound recognition. This tool was built as a Python script based on the library Persephone [10]. This library is meant as a speech recognition tool for transcription of low-resource languages and contains several parts (see Fig. 1).

[3] The explanatory notes listed below are only the ones relevant for the listed example and are not the complete set.

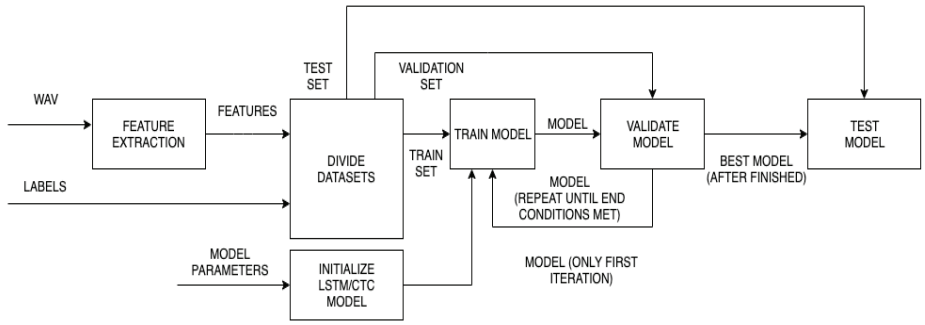


Fig. 1. Individual speech sound recognition tool architecture

The first part is audio feature extraction tools, for our experiment we have used LMFB (Log Mel Filterbank) with delta and delta-delta features. After we extracted features from *wav* recordings, we used Persephone functions to split data (features + labels) into three non-intersecting sets in the following way: 90% of data to train set, 5% to validation set and 5% to test set. We initialised our model when we had prepared our data. "The underlying model used is a long short-term memory (LSTM) recurrent neural network [11] in a bidirectional configuration [12]. The network is trained with the connection's temporal classification (CTC) loss function [13]." [10] We have used default three-layered architecture with 250 hidden nodes. This model is then trained with pre-processed data for at least 30 epochs. Training stops when one of the following conditions is met:

(a) training LER (learning error rate) is lower than 0.1% and the validation LER is lower than 1%;

(b) validation LER has not improved in the last 10 epochs;

(c) after 100 epochs.

In the last step, we tested our trained model against the test data set.

This tool was initially tested on tonal languages and thus it provides the ability to label prosodic features such as tone or word stress. We decided, however, not to use those features as individual labels in the first version of our annotation system. The tool is designed to transcribe whole utterances, however, in our case our utterances are only individual sounds so the label always corresponds to a single speech sound.

For our experiment, we had 3,717 labelled sounds from 32 non-native Czech speakers. When we tried to train the tool with data labelled with the initial version of the annotation, the model stopped after 57 epochs with a huge training error rate 43.4% and a validation rate of 42.4% and an even worse error rate of 50.8% for the

test set. After checking the model results on the test set, however, we found some interesting data. Most of the incorrectly labelled data were consonants and even in those cases, the model output was often a consonant that differed from the expected consonant only in several features such as voicing, articulation position, fricative vs affricate or different variants (aspirated *t* vs "hard" *t*) as shown in Tab. 1.

Table 1 also shows one very interesting case of mislabelling that unveiled one of the issues with annotation. In the last line of Table 1, it is apparent that the expected label was *z::dz* which means that *dz* was pronounced by the speaker instead of *z* and the output label is *dz*. The problem with this is that both of those labels correspond to the same pronunciation, which is *dz*. This leads to an update in our annotation: we have changed the annotation of the incorrect sound from format X::Y (X being the expected and Y being the pronounced sound), to simply Y.

expected label	**output label**
h [ɦ]	ch [x]
s	z
m	n
ch [x]	f
c [t͡ʂ]	dz
z	dz
g	d
t:vA	t:vT
z::dz	dz

Tab. 1. Expected vs output label (consonants)

As can be seen in Table 2, the vowels were in most cases annotated correctly. There were no issues in vowel quality except for cases when diphthongs were classified as simple vowels. This happened especially for diphthongs with a shortened second component or diphthongs and lengthened vowels. There were also some issues with quantity identification. We also observed the same problem with duplicate annotation of format X::Y vs Y.

expected label	**output label**
eu:kK_2 (shortened *u*)	e [ɛ]
eu:vZ_1 (closed *e* [ɛ])	e:vZ (closed *e* [ɛ])
u:kD (lengthened)	u
au [a͡u]	a:kD (lengthened)
eu [ɛ͡u]	e [ɛ]
ou:kK_2 (shortened *u*)	ou [o͡u]
i:kD (lengthened)	i [ɪ]
au::a	a

Tab. 2. Expected vs output label (vowels)

Cases where consonants were annotated as vowels or vice versa were exceptionally rare and for an undiscovered reason. the most frequent error of this kind was labelling *p* as *e* [ɛ]; this could be possibly due to some mistake in the annotation.

After these findings, we decided to go through the annotation and attempt to identify duplicate labels such as the mentioned X::Y vs Y case. By unifying those labels, corresponding to the same sound, and removing a few less important features, we have dramatically reduced the label inventory size, which led to much better results.

The adjusted annotation model stopped after 70 epochs with a much lower training error rate 14.9% and validation error rate 36.8%. The test error rate also improved to 41.27%. Putting aside X::Y vs Y case, errors in the output of the new model were similar to the previous one, although they were less frequent.

There are two main consequences of those results. It is apparent that although we had quite a small data set and the model was far from an optimized one, we ended up with quite good results, although they are still not good enough to be used in a real-world application. The second one is that label inventory size has a huge impact on success rate (along with the amount of available data) and that we have to avoid different labels for the same or very similar sounds at any cost.

6 FUTURE PROSPECTS

The findings from the first testing of our approach lead us to several ideas on how to improve our system. The first plan in the future is to split the annotation into two parts. The first part would be the labels corresponding to each of the individual segments. This would be a simple identifier in the form of a number or character string. These segments would be annotated as *1* or *a1* instead of, for example, *a:kD*. The second part of the annotation will be a mapping table that will translate the identifier to its corresponding attributive annotation that will be used to obtain feedback based on speech recognition output. In this part we would consequently have the information that *1* corresponds to *a:kD*. We also want to try to split certain features such as length, nasalisation, aspiration or stress to individual segments, thus instead of *á* we would have *a:* and instead of *p:vA* (aspiration) we would have *p>* (where > means the aspiration segment). This will allow us to easily extend our annotation, shrink the size of our label dictionary, and focus on the most frequent non-native sounds in Czech.

In conclusion, the field of speech recognition in e-learning and automatic feedback on non-native speech is still in its beginnings, but our findings could become the basis for a new approach to this complex and increasingly important problem. A great deal of this research, however, still has to be done and much data has to be collected to create a system that can be used in e-learning systems. Non-native speech recognition is nevertheless a topic to be considered.

ACKNOWLEDGEMENTS

The research was supported by the Ministry of Education of the Czech Republic IGA_FF_2020_021 "Czech Studies: Literary and Linguistic Overlaps and Interpretations" and MUNI/IGA/1225/2020 "L2 Pronunciation system (Linguistic Annotation System)".

References

[1] Blake, R. (2011). Current Trends in Online Language Learning. Annual Review of Applied Linguistics, 31, pages 19–35.

[2] Duolingo, Inc. (2021). Duolingo (5.1.5) [Mobile app]. Accessible at: https://play.google.com/store/apps/details?id=com.duolingo.

[3] Mikušiak, L. (2015). Ten Ta To (1.7) [Mobile app]. Accessible at: https://play.google.com/store/apps/details?id=com.lubosmikusiak.articuli.tentato.

[4] EVE Technologies, s.r.o. (2021). CzechME (1.0.5) [Mobile app]. Accessible at: https://play.google.com/store/apps/details?id=cz.evetech.czechme.

[5] Rootstrap, Inc. (2020). Report: Online Education Industry Growth 2020. Rootstrap [Web page]. Accessible at: https://www.rootstrap.com/annual-report-online-education-statistics.

[6] Kuhn, R., and De Mori, R. (1990). Cache-based natural language model for speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(6), pages 570–583.

[7] ELSA Co, Ltd. (2021). ELSA Speak: Online English Learning & Practice App (6.2.1) [Mobile app]. Accessible at: https://play.google.com/store/apps/details?id=us.nobarriers.elsa.

[8] Becker, K., and Edalatishams, I. (2019). ELSA Speak – Accent Reduction [Review]. In J. Levis, C. Nagle and E. Todey (eds.), Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference, Ames, IA: Iowa State University.

[9] Hedbávná, B., Janoušková, J., and Veroňková, J. (2009). Výslovnost češtiny u cizinců – poznámky k metodám výuky. In Sborník Asociace učitelů češtiny jako cizího jazyka 2007–2009, pages 16–23, Praha: Akropolis.

[10] Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S. et al. (2019). Evaluating phonemic transcription of low-resource tonal languages for language documentation. LREC 2018 (Language Resources and Evaluation Conference), pages 3356–3365, Miyazaki, Japan.

[11] Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), pages 1735–1780.

[12] Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), pages 2673–2681.

[13] Graves, A., Fernandez, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. Proceedings of the 23rd international conference on Machine Learning, pages 369–376.

sciendo
DOI 10.2478/jazcas-2021-0048

DESIGNING A CORPUS OF CZECH MONOLOGUES: ORATOR v2

MARIE KOPŘIVOVÁ – ZUZANA LAUBEOVÁ – DAVID LUKEŠ
Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic

KOPŘIVOVÁ, Marie – LAUBEOVÁ, Zuzana – LUKEŠ, David: Designing a corpus of Czech monologues: ORATOR v2. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 520 – 530.

Abstract: ORATOR v2 is a new 1.5M word corpus of Czech monologues, delivered to a live audience in semi-formal to formal settings. It was designed to chart the space of naturally occurring monologues which can be obtained for corpus processing. As such, it aims for diversity but does not attempt any balancing of subcategories, recognizing that some types of data are inherently easier to obtain in high volume than others. The transcription guidelines and annotation tools employed are the same as other recent spoken corpora published by the CNC, which facilitates interesting comparisons between various types of spoken Czech. The present paper sketches out three case studies, comparing ORATOR to the informal conversations of ORTOFON v2 in terms of the frequencies of demonstratives and hesitations, as well as lexical richness.

Keywords: speech, corpus, monologue, Czech

1 INTRODUCTION

With regard to spoken language, the Czech National Corpus (CNC) has historically mainly focused on collecting recordings of multi-party conversations in an informal setting, among friends and family. These interactions are thematically unspecified and unprepared; throughout the years, they have been made available to the public in a long line of corpora, culminating in the ORTOFON corpus, whose version 2 was published at the end of 2020 [1].

At the same time, after a preliminary version 1 in 2019, the full version 2 of the ORATOR corpus was also released [2]. ORATOR marks a departure from the relatively narrow focus on informal spoken Czech: it contains recordings and transcripts of mostly semi-prepared monologues of various kinds, providing a window to the opposite side of the spectrum of spoken communication. Since both corpora adhere to the same transcription guidelines[1] and were lemmatized and morphologically tagged using the same system [3], we hope they will not only enable a wealth of comparative research into various registers of spoken Czech, but they will also make interpretation of the results exceptionally straightforward and reliable.

[1] Except for the phonetic transcription layer, which is absent in ORATOR.

2 MONOLOGUE: DEFINITION AND THE EXISTING CZECH CORPORA

Spoken communication is traditionally divided into monologue and dialogue. This kind of classification is based on the number of active speakers (subjects): one only establishes a monologue, more than one a dialogue. Hoffmannová [4] defines monologue as an uninterrupted continuous activity of one subject and points out that pure monologues are very rare. Monologue is always more or less dialogical, depending on the degree of focus on the recipient, and the same is true vice versa.

Within monologues, many different genres or text types can be distinguished. Müllerová [5] introduces e.g., the following: narration of a story or memories (often with description of places or persons), introduction to a discussion, lecture, ceremonial official speech, sermons, etc.

Of course, the ORATOR corpus is hardly the first corpus of Czech to include monologues. The first spoken corpus within the CNC project – the Prague Spoken Corpus (PSC) [6] – combines two types of documents: informal, unprepared dialogue, and a structured interview with open questions. In response to the questions, speakers usually produced extensive monologues, as befits an interview. The Brno Spoken Corpus (BSC) [7] has a similar design. The formalization of the question–answer sequence led to these parts of the PSC and BSC being branded as "formal". It is however a slightly different type of formality than that in the ORATOR corpus (see below for details).

A corpus which consists entirely of pure monologues is the (aptly named) MONOLOG corpus [8]. The recordings feature a prepared and mainly read out speech by professional speakers of the Czech Radio.

3 CRITERIA FOR INCLUSION

Compared to the above-mentioned corpora, what makes ORATOR stand out is its strong emphasis on collecting naturally occurring semi-prepared monologues, e.g., university lectures, as opposed to ones that are experimentally induced and/or fully read out. Many spoken corpora focus on recordings of lectures and seminars for pragmatic reasons, because of their relative obtainability and consistent quality, which makes them well-suited for automated processing and use in NLP or ASR. By contrast, the ORATOR corpus has a broader scope: it was created as an intentional exploration of the different types of monologues which occur in communication.

Data collection focused on communication situations which are specifically intended to stand on their own as monologues. These may later be followed by a dialogical part (e.g., a discussion after a lecture) which is, however, not included. Some monologues may be part of more complex situations, such as meetings. Nevertheless, it is always the case that one speaker speaks without interruption,

having been allocated space and time for his or her speech, and the monologue and the dialogical part are separate. Also included were sequences of monologues linked by a moderator's commentary, such as introductory speeches at the opening of an exhibition, as well as less typical monologues, such as yoga classes or workplace fire safety trainings.

No balancing criteria were set in advance, the aim was simply to create the most diverse corpus of monologues possible and to find out which types can be obtained. Certain types of communication cannot be made public for legal or ethical reasons, and some are not appropriate because they formally intertwine short spans of monologue and dialogue in such a way that disentangling them would make the entire structure collapse.

The following criteria (cf. [9] for more details) for inclusion of a candidate recording in the corpus were determined:

1. A self-contained stretch of **monologue** by a speaker who is informed in advance about the topic, occasion, time, and location of his speech. The speaker can use different levels of preparation, such as notes, projected presentations, photographs, etc. We originally excluded speeches which were entirely or partially read out. However, this would deprive us of some types of situations, part of which requires a precisely given form, for instance because it is also a legal act (e.g., a wedding ceremony). In the case of lectures, they can contain quotations which are usually read out. The preparation of a text intended for reading also has its specificities, which is why we ended up including a small minority of these recordings to complete the picture (18 in total).
2. The context can be described as **official**, at least to a degree, and speakers were appointed either due to their expertise (public lecture, professional training, etc.), institutional role (university lecture, mayor's speech, etc.) or social status within the group (e.g., during a wedding toast). In some cases, the asymmetry of communicative roles was strengthened by the presence of a moderator. However, in smaller groups, this difference was weakened, sometimes questions were asked during the speech, especially in training sessions.
3. **Liveness**: we selected situations in which the speaker addresses a group of listeners. These ranged from smaller professional or private groups (training, wedding), to larger communities (lectures, sermons) to completely public speeches (public gatherings). We mostly wanted to avoid pre-recorded speeches which could be edited or adjusted for a particular platform. Still, as in the case of reading, we broke this rule in a few cases (9 monologues recorded specifically for an internet audience) for the sake of diversity.

The speakers were consistently anonymized and no sociolinguistic categories were identified, apart from gender. Gender information was also used to generate

nicknames for the speakers, based on randomly selected surnames supplemented by a first name initial, e.g. *Tománková, O.*[2]

4 OVERVIEW AND STATISTICS

ORATOR v2 contains over 1.5M positions in 489 recordings from 2005–2019 by 468 different speakers (some short speeches connected by the moderator form a single document and, conversely, some long lectures are divided into several parts). The ratio of recordings made specifically for this corpus vs. those acquired from external sources is about 4:3, and their length ranges from 13 seconds to 49 minutes, for a total length just shy of 149 hours. Men dominate significantly in the corpus, accounting for 71% of the number of tokens and 69% of the speakers.

As for document-level metadata, we tried to provide multiple grouping perspectives, so as to help users find their way around the corpus. Firstly, the **situational frame**: speeches were divided into official (at exhibitions, graduations, wedding ceremonies), popularizing (lectures for the public), political, professional (training sessions) and scientific (university and conference lectures). Table 1 shows the number of positions and documents in the corpus broken down by frame. Clearly the official recordings, while relatively numerous, are mostly quite short, as can be expected from the examples above. In the popularizing, professional, and scientific frames, longer lectures dominate, accounting for 49% of recordings and 78% of positions.

Secondly, 12 **situation types** provide a more fine-grained categorization of the recordings. A breakdown with examples is given in Table 2.

Thirdly, **genre** was annotated following the categories used in the latest SYN series written corpora, starting with SYN2015 [10]. While not all categories are represented, the sample is still varied and allows for interesting comparisons between written and spoken texts within a given genre.

These main divisions are complemented with information about the intended audience (public vs. restricted) and a special field identifying fringe types of monologue which technically did not meet criteria for inclusion, but were included in small amounts for diversity (cf. some examples in Section 3).

5 COMPARISON WITH ORTOFON V2

In many ways, the monologues in ORATOR are a stepping stone between the spontaneity of informal dialogues and the level of preparedness of written texts. We

[2] This is intended to remind the corpus user that the recordings were made in relatively formal/public settings. By contrast, speakers featured in the private conversations of ORTOFON v2 are identified by randomly selected first names with a surname initial, e.g., *Aleš N.*

are, therefore, convinced they will form an interesting basis for comparative research. Three simple case studies are presented in the following subsections to give an idea of the possibilities: comparisons of demonstratives, hesitations, and lexical richness between ORATOR v2 and the ORTOFON v2 corpus, where the latter consists of informal conversations. We hypothesized there would be more demonstratives in ORTOFON (as conversations are more heavily context-embedded), more hesitations in ORATOR (speakers tend to avoid long stretches of silence in monologues, leading to a higher incidence of filled pauses), and higher lexical richness in ORATOR (since the monologues are mainly expository and information-heavy).

5.1 Demonstratives

The relative frequency of demonstratives[3] in ORATOR and ORTOFON is given in Table 3. It shows that demonstratives are slightly (1.2×) more common in ORTOFON, i.e., in informal spontaneous conversations. The most frequent demonstrative lemma is *ten* 'this', which covers 92% of all demonstrative occurrences in ORTOFON, but only 86% in ORATOR (though still at the top of the frequency list).

This raises the question, where did those 6 percentage points get redistributed to? Part of the answer might be towards "long" demonstratives[4] such as *takovýhle* 'such a one' or *tenhleten* 'this one'. As Table 3 indicates, with these, the situation is reversed: they are actually about 1.3× more common in ORATOR.

In this light, our conjecture that dialogues contain more demonstratives because of their context-embeddedness might need revisiting. "Long" demonstratives, reinforced by the use of morphemes such as *hle*, are actually the ones which retain strong semantics of co(n)textual reference, and might be motivated by the frequent use of props such as photographs or slides during monologues. By contrast, the most frequent word form in ORTOFON overall is *to*, which is formally part of the paradigm of the demonstrative *ten*, but often performs more of a connective function, especially when switching speakers in dialogue (*to jo* 'yes'). As there is no speaker switching in ORATOR, the frequency list is topped instead by *a* 'and', another connective, which is arguably more useful in monologues (it is also typically the most frequent word in corpora of written Czech). So the difference in the incidence of demonstratives between monologues and dialogues might have more to do with different discourse structuring patterns rather than varying levels of context-embeddedness.

5.2 Hesitations

The transcription of both corpora notes certain non-verbal sounds, including hesitations or filled pauses, transcribed as @ (short) or @@ (long). Functionally,

[3] Retrieved via the query `[tag="PD.*"]`.

[4] Retrieved via the query `[tag="PD.*" & word=".{5,}"]`.

hesitations are connectives: speakers use them to eliminate (silent) pauses and fill the time needed to think their next utterance through [11]. As for listeners, they tend to perceive them negatively, as parasitic filler sounds [12].

Looking at the comparison in Table 4, hesitations of both types are clearly much more common in monologues. In both corpora, they tend to co-occur with pauses and other connectives such as *a* 'and' or *že* 'that'. In fact, when combined, they dominate the category of linking devices in ORATOR by a large margin, at 30,393 i.p.m.: the most common word in this category is the conjunction *a*, at 25,824 i.p.m.

Hesitations often appear when speakers attempt to convey complex notions, struggle finding the right word, or experience stress and/or high cognitive load, which would explain their increased presence in formally constrained monologues as opposed to freeform informal conversations, especially since the proportion of hesitations is highest among lectures, especially scientific ones. Intriguingly enough, they also appear in read-out speech, though at a relatively low frequency.

The lowest overall frequencies, even lower than ORTOFON, are encountered in sermons and ceremonies, though it should be noted that as with read-out speech, these are small categories with little data, so any generalizations are tentative at best. Still, a possible cause might be that speakers try to conform to a higher standard on these occasions, or that they occur repeatedly, leading to a high degree of preparation, or perhaps individual speaker proficiency.

5.3 Lexical richness

Finally, we turn to lexical richness. A naive measure of lexical richness is the type-token ratio (TTR), which is, however, sensitive to text length, as is well-known. Therefore, we used two more sophisticated TTR-based measures: the moving-average TTR (MATTR) [13] and zTTR[5] [14]. While MATTR is calculated by sliding a fixed-size window over the text and averaging the obtained TTR values, zTTR gives the relative position of a text within a reference distribution of texts of similar length.

We ran data extraction under a variety of settings:

- window sizes of 100 or 500 tokens for MATTR
- journalistic texts or spoken dialogues as reference data for zTTR
- tallying word forms or lemmas
- per document or speaker in the document

The general shape of the results was similar across the board, so we only selected three representative examples (Figures 1–3), all word-form based, subdivided by corpus (ORATOR vs. ORTOFON) and target unit (document vs.

[5] We are grateful to Václav Cvrček for letting us use his Perl script and reference data to calculate zTTR values.

speaker in document). The left subplots show median MATTR/zTTR values with bootstrapped 99% confidence intervals computed via 10,000 iterations of Monte Carlo case resampling; the right subplots show probability density functions for the full distribution of MATTR/zTTR values in each corpus, computed via kernel density estimation.

First and foremost, what all figures clearly show is that ORATOR monologues tend to have higher lexical richness than ORTOFON dialogues, whichever way we slice them (by document or speaker). This is consistent with our expectations, based on the fact that the monologues are mostly expository – speakers are primarily trying to convey information and have a limited timeframe to do so. This makes them aim for information-dense speech, which favors increased lexical richness.

Another observation is that in the case of ORATOR, there is little difference when calculating TTR per document vs. per speaker: the density curves and confidence intervals for the medians are nearly identical, or at least overlap to a great extent (Figure 3). This makes sense: in ORATOR, documents mostly feature a single speaker, so there is little difference in the units for which TTR is calculated to begin with.

In the case of ORTOFON however, the per-speaker distributions are consistently shifted to the right, towards (slightly) greater lexical richness: a little in the case of MATTR in Figure 1 (though note that the confidence intervals for the medians do not overlap, so this looks like a reliable effect, though small), and some more with zTTR, especially in Figure 3. Since ORTOFON documents are dialogues, slicing them up by speaker actually does make a difference in the units, but the fact that it does have an effect on TTR was still somewhat surprising to us. It remains to be seen whether an underlying linguistic explanation can be uncovered, or whether this is a residual failure of both measures to compensate for different text lengths.

Turning our attention to Figure 2, we see that the per-speaker density curve for ORTOFON peaks at 0, which is a good sanity check: it shows that our sample has the same mean as the reference data extracted from another corpus of informal dialogues. The ORATOR distributions peaking above 0 is a further confirmation of the fact that semi-prepared monologues tend to be lexically richer than informal dialogues.

Finally, the journalistic zTTR was included because journalistic texts spread over a large, well-populated portion of the TTR landscape, which makes them useful as reference data for comparison across registers. Figure 3 shows that in general, the lexical richness of both dialogues and monologues is on the low end of the spectrum, with all four distribution curves squeezing almost entirely below 0, i.e., the average.

6 CONCLUSION

The ORATOR v2 corpus is freely available via the KonText search interface at https://korpus.cz/kontext; other types of access to the data can also be provided upon

request.[6] As we have sketched above, it presents many compelling research opportunities: fruitful comparisons can be drawn both within the corpus itself and with other corpora. ORTOFON v2 is an especially attractive option in this regard because the two corpora focus on opposing ends of the spoken Czech spectrum while sharing the same processing pipeline, which makes it less likely for researchers to be misled by spurious differences caused by arbitrary incompatibilities between corpora. In the previous sections, we have given a glimpse of the possible directions to explore, but these are obviously just a tip of the iceberg. We are looking forward to see what creative uses these resources will be put to by fellow linguists.

ACKNOWLEDGEMENTS

The design and compilation of CNC corpora is made possible by the Czech National Corpus project (LM2018137) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation.

Frame	Positions	Documents
popularizing	812,671	188
scientific	361,770	75
professional	178,229	52
official	164,013	164
political	18,966	10

Tab. 1. Number of positions and documents in ORATOR v2, broken down by situational frame

Situation type	Positions	Documents
lecture (academic, general public)	1,204,668	240
public assembly	63,891	24
meeting	49,451	19
tour (e.g. castle tour)	43,073	31
opening speech	34,644	64
introduction of a work of art	33,931	35
training (e.g. workplace safety)	32,438	11
instructions (e.g. yoga class)	26,176	12
celebratory address	17,483	20
ceremony (e.g. wedding)	12,658	14
sermon	9,087	8
closing speech	8,149	11

Tab. 2. Number of positions and documents in ORATOR v2, broken down by situation types

[6] Please use the form at https://korpus.cz/clarin/helpdesk to submit your request.

Type of demonstratives	i.p.m. in ORATOR v2	i.p.m. in ORTOFON v2
all	65,156.17	78,221.82
"long"	7,782.38	6,025.17

Tab. 3. Comparison of the relative frequency of different types of demonstratives (in instances per million)

(Sub)corpus	i.p.m. of @	i.p.m. of @@
ORTOFON v2	7,613	1,606
ORATOR v2	24,797	5,596
- read	5,939	2,124
- lectures	26,284	5,890
- scientific	35,329	8,986
- sermons	6,933	550
- ceremonies	6,320	1,185

Tab. 4. Relative frequency of short @ and long @@ hesitations in various (sub)corpora (in instances per million)

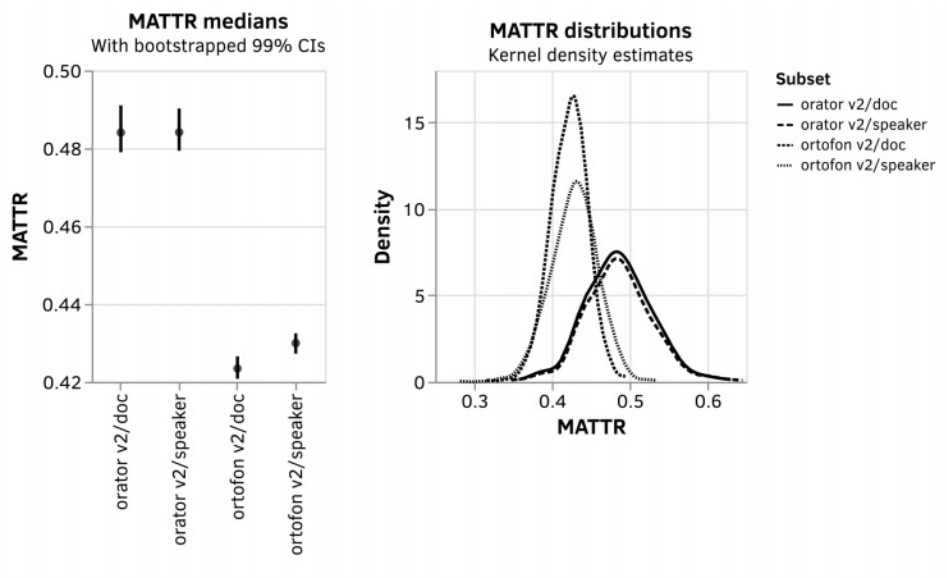


Fig. 1. MATTR computed on word forms, with a window of 500 tokens

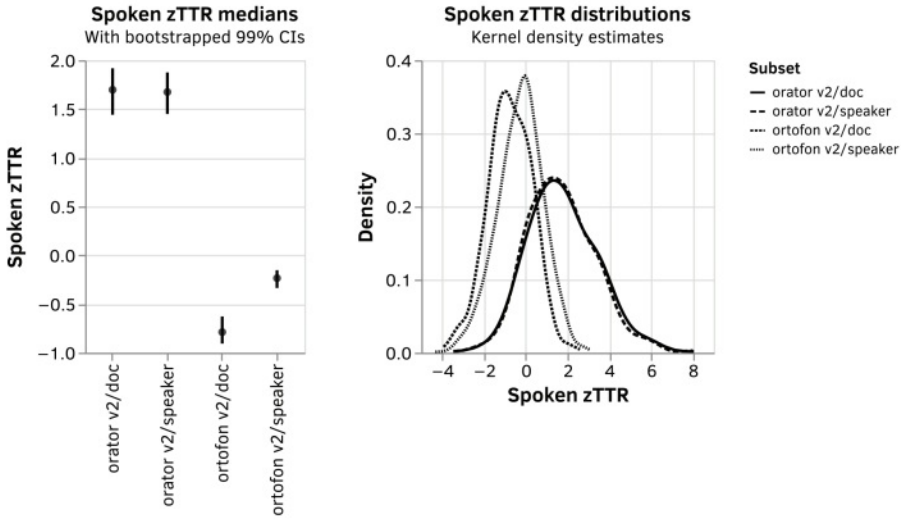


Fig. 2. zTTR computed on word forms, against spoken dialogue reference data

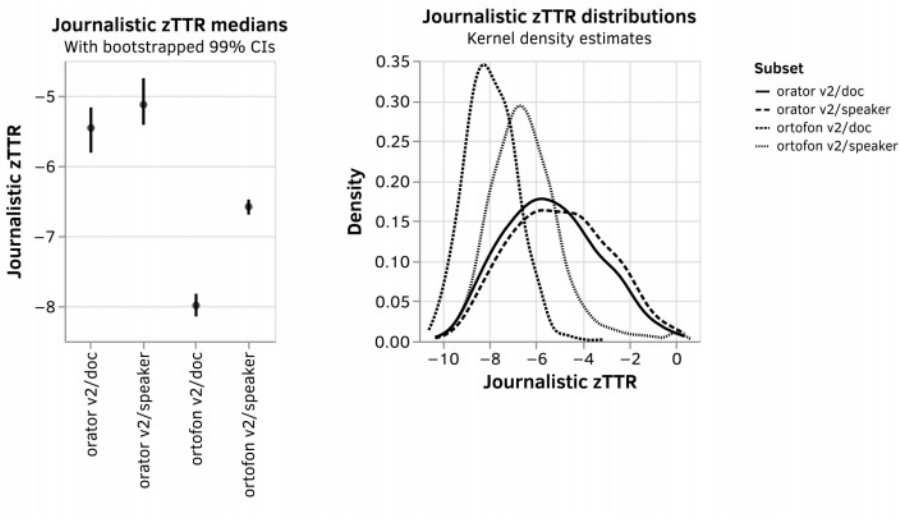


Fig. 3. zTTR computed on word forms, against journalistic reference data

References

[1] Kopřivová, M., Laubeová, Z., Lukeš, D., Poukarová, P., and Škarpová, M. (2020). ORTOFON v2: Korpus neformální mluvené češtiny s víceúrovňovým přepisem. ÚČNK FF UK: Prague. Accessible at: https://korpus.cz.

[2] Kopřivová, M., Laubeová, Z., Lukeš, D., and Poukarová, P. (2020). ORATOR v2: Korpus monologů. ÚČNK FF UK: Prague. Accessible at: https://korpus.cz.

[3] Kopřivová, M., Komrsková, Z., Lukeš, D., and Poukarová, P. (2017). Korpus ORAL: sestavení, lemmatizace a morfologické značkování. Korpus – Gramatika – Axiologie, 15, pages 47–67.

[4] Hoffmannová, J. (2017). Monolog. CzechEncy – Nový encyklopedický slovník češtiny. Accessible at: https://www.czechency.org.

[5] Müllerová, O. (2000). Žánry a syntaktické rysy mluvených projevů. In Tváře češtiny, pages 21–54, Ostrava. Ostravská univerzita.

[6] Čermák, F., Adamovičová, A., and Pešička, J. (2001). PMK: Pražský mluvený korpus. ÚČNK FF UK, Praha.

[7] Hladká, Z. (2002). BMK: Brněnský mluvený korpus. ÚČNK FF UK, Praha.

[8] Štěpánová, V. (2016). Korpus Monolog 1.1. Accessible at: http://monolog.dialogy.org.

[9] Kopřivová, M., Komrsková, Z., Poukarová, P., and Lukeš, D. (2019). Relevant criteria for selection of spoken data: theory meets practice. Jazykovedný časopis, 70(2), pages 324–335.

[10] Křen, M. et al. (2016). SYN2015: Representative Corpus of Contemporary Written Czech. In Proceedings of LREC, pages 2522–2528, Portoroz. ELRA.

[11] Čermáková, A., Jílková, L., Komrsková, Z., Kopřivová, M., and Poukarová, P. (2019). Diskurzní markery. In Syntax mluvené češtiny, pages 244–351, Prague. Academia.

[12] Skarnitzl, R., and Machač, P. (2012). Míra rušivosti parazitních zvuků v řeči mediálních mluvčích. Naše řeč, 95, pages 3–14.

[13] Kubát, M., and Milička, J. (2013). Vocabulary Richness Measure in Genres. Journal of Quantitative Linguistics, 20(4), pages 339–349.

[14] Cvrček, V., and Chlumská, L. (2015). Simplification in translated Czech: a new approach to type-token ratio. Russian Linguistics, 39(3), pages 309–325.

sciendo
DOI 10.2478/jazcas-2021-0049

SHARING DATA THROUGH SPECIALIZED CORPUS-BASED TOOLS: THE CASE OF GramatiKat

DOMINIKA KOVÁŘÍKOVÁ
Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic

KOVÁŘÍKOVÁ, Dominika: Sharing data through specialized corpus-based tools: The case of GramatiKat. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 531 – 544.

Abstract: This paper presents a specialized corpus tool GramatiKat in the context of Open Science principles, namely data sharing, which offers opportunities for original research and facilitates verifiability of research and building on previous research. The tool is designed primarily for examining grammatical categories from the quantitative point of view. It offers grammatical profiles of particular lemmas (currently 14 thousand Czech nouns) and the proportion of individual grammatical categories within a part of speech, i.e., the standard behavior of a word class. The data in GramatiKat are pre-processed, statistically evaluated, and presented in charts and tables for clarity, and they are available to other linguists, especially from fields of morphology and lexicography. This article is aimed at providing inspiration and support to corpus and non-corpus linguists with utilization and enhanced use of the existing tools and with the creation of new specialized tools available to other users.

Keywords: specialized corpus tools, grammatical category, morphology, lexicography, Open Science

1 CORPUS LINGUISTICS IN THE CONTEXT OF OPEN SCIENCE

Current trends of open access to research outputs and of data sharing, which are among the principles of Open Science, are key themes of the contemporary research community. In corpus linguistics, this is not a new topic; corpora themselves, as well as corpus concordancers, are research outputs that allow both the verifiability of research conducted on corpus data and the building on previous research, while offering all users vast opportunities for original research in various fields of linguistics. These are precisely the requirements formulated by J. Chromý and V. Cvrček [1, pp. 8–11] in the article opening the monothematic issue of *Naše řeč* (1/2021), which set itself the task of opening a broad discussion on the topic of open linguistics. Contributions to this discussion range from appeals and program statements, to organizing projects aiming at data sharing, and to the actual implementation of the principles in the form of shared research articles, data, software, or other tools. I would like to contribute as well, specifically in the area of "synergy and cooperation between the researchers" [1, p. 5].

Just ten years ago, in 2011, an article was published in *Naše řeč*, that was reflected on in an editorial of the *Jazykovedný časopis* in 2019 (2/2019). The article "Možnosti a meze korpusové lingvistiky" [2] focuses, among other things, on changing trends in corpus linguistics, a discipline that adopted the principles of sharing data and tools for their analysis from the very beginning of its existence. In the first 20 years of its greatest boom since the late 1980s, corpus linguistics was devoted first to data collection and tagging, and subsequently to diverse and extensive linguistic research enabled by high-quality, large-scale data and to expanding possibilities for analysis.

In the 10 years that have passed since the 2011 article, another strong trend can be observed: development of specialized tools to process corpus data. Such tools facilitate data analysis methods, such as keyword analysis or statistical evaluation of corpus data, or they offer pre-processed data to enable research focused on particular areas (e.g., identifying metaphors or phraseology, finding n-grams, exploring translation equivalents or vocabulary of a particular text type). This trend is noticeable in corpus linguistics worldwide (see webpage providing links to various corpus tools https://corpus-analysis.com/) and the principles of Open Science have been incorporated in the Czech National Corpus project as well. Seven publicly available tools have been published in the last three years alone, offering statistical data analysis, pre-processed and organized datasets or data visualization in the form of interactive tables, graphs, and dialectological maps.[1]

As a co-author of two tools aimed at assisting other linguists with examining research areas of grammatical categories and academic vocabulary (both with Oleg Kovářík), I would like to share my experience with the development and application of such tools (specifically, I will focus on the GramatiKat tool [3]). Hopefully, this article will provide inspiration and support to corpus and non-corpus linguists in utilizing and enhancing the existing tools, sharing ideas and resources such as access to data or programming skills, and provide other researchers with new tools and pre-processed data for their original research.

2 GramatiKat: TOOL FOR RESEARCH OF GRAMATICAL CATEGORIES

The GramatiKat tool is designed primarily for researching grammatical categories in Czech. The idea of examining grammatical categories from a less traditional, quantitative point of view originated many years ago, during work on the corpus-based *Mluvnice současné češtiny* [16, pp. 205–209], and was sparked by

[1] Tools created within the Czech National Corpus project: SyD [4], Morfio [5], KWords [6], Treq [7], Pro školy [8], Slovo v kostce [9], Calc [10], Lists [11], KorpusDB [12], QuitaUp [13], Mapka [14], Akalex [15] and GramatiKat [3]. Manuals to and information on all the tools are available at https://wiki.korpus.cz/doku.php/en:manualy.

research on gradation of adjectives. This phenomenon stands between grammar and word formation, partly because it does not apply to all adjectives. In fact, we found out that comparative and superlative forms are attested only in a fraction of adjectives – in the most recent corpus of contemporary written texts SYN2020 [17], only about 10% of adjectives (with frequency 3 or more) have comparative or superlative form, i.e., less than 4 thousand adjectives. Among them, however, there are adjectives with a very high frequency, so graded forms are encountered quite often in texts, and gradation is considered a relatively common phenomenon.

The primary goal of GramatiKat is to expand this initial idea of quantitative research to all grammatical categories in all parts of speech, especially nouns, adjectives, and verbs. Such information is not accessible through a standard corpus concordancer search and so only someone with special resources (access to data and programming skills) is usually able to carry out such research. Through the GramatiKat tool, the data are available to all interested researchers. It would be, of course, possible to share the raw data with a presumption that experienced users can draw their own conclusions. However, we have chosen a more involved approach. The data is pre-processed, statistically evaluated, and presented in charts and tables (and of course, the raw data is also available).

The first version of the tool has been available since early 2021 and includes information about Czech nouns and their categories of number, case, and gender. For the Slovko 2021 conference, data for Slovak nouns were added (see more in section 4.4).

The information that is currently available in the GramatiKat tool includes:

- distribution of grammatical category values within a word class, e.g., distribution of all 14 cases (7 cases in singular and plural) in Czech nouns. For example, a chart (identical to figure 1 in section 4.1) shows the percentage of locative singular or dative plural;
- distribution of grammatical category values within a lemma, or, a grammatical profile [18, p. 11], e.g., what is the grammatical profile of a lemma *večer* 'evening';
- a list of words that show an unusually high frequency of a grammatical category value, e.g., individual nouns that are attested significantly more often than other nouns in a specific case;
- a list of words with a gap (or unattested form) in the paradigm, e.g., singularia tantum.

3 MATERIAL AND METHODS

We used data from the SYN2015 [19] representative corpus of contemporary written Czech with 120 million words (incl. punctuation) to create the GramatiKat tool. The corpus is balanced and consists of 1/3 fiction, 1/3 journalistic, and 1/3 non-

fiction and academic texts. In the first version of GramatiKat, we included all nouns from the SYN2015 corpus with a frequency of at least 100[2] (14 thousand noun lemmas).

For comparison of Czech and Slovak nouns, we prepared a special parallel subcorpus of InterCorp version 13, containing parallel Czech and Slovak texts. The subcorpora size is 50 million lines (incl. punctuation) in Czech, 49 million lines in Slovak. For Czech, we examined 5600 lemmas with a frequency of at least 100, for Slovak 5400 lemmas.

We examined three grammatical categories, or rather three combinations of grammatical categories: the number, the combination of case and number (i.e., 14 paradigm cells), and the combination of case and number with gender.[3]

For statistical evaluation of the data, a boxplot was used. In addition to its usual purpose, which is visualization of numerical data in quartiles, a boxplot can also serve as a guide to estimate which values are standard and which are exceptions, in other words, which values are unusually high or unusually low. Such outliers are often calculated as exceeding 1.5 times the interquartile range above the third quartile and below the first quartile (although there are other options for evaluation, e.g., using standard deviation; outliers can also be disregarded altogether).

The boxplots in GramatiKat not only show but also determine the standard and non-standard behavior of the whole part of speech. In the context of the presented research, we consider the values that do not belong to the outliers to be the standard behavior of Czech nouns in a given case, and outliers from 1.5 times above the third quartile to be exceptions – words with an unusually high representation of the given case. The lower outliers are not present in our data at all (the 1.5 times the IQR below the first quartile reach zero or negative values in all cases), and we consider the absence of a certain form in a corpus (or, a gap in the paradigm) to be non-standard behavior.[4]

In examining the quantitative properties of grammatical categories, it is necessary to be aware that the percentage of values in each grammatical category can be influenced by various factors, particularly by the size and composition of the corpus and the frequency of the lemma. A large representative and balanced corpus with a wide range of text types in balanced proportions such as SYN2015 ensures a high degree of reliability of the grammatical profiles, at least within the language

[2] We have chosen a relatively high frequency so that the probability of a given form would be high enough.

[3] In the future, we intend to include other parts of speech, primarily adjectives and verbs. In addition to traditional grammatical categories, we would like to thoroughly examine negation, which has not yet received sufficient attention in Czech grammatical or lexicographic descriptions or even corpus lemmatization (adjectives).

[4] The vocative is an exception: both singular and plural are usually unattested, so the gap in the paradigm is actually standard behavior.

variety under consideration. The researcher should always be aware of this limitation and especially of the influence a smaller or unbalanced corpus may have on the results (see sections 4.3 and 4.4).

4 RESULTS

4.1 Distribution of a grammatical category of case in Czech nouns

The main information that the user can get from the GramatiKat tool is the overview of a given grammatical category within a certain part of speech. Figure 1 gives an overview of the case (in combination with number[5]) distribution in Czech nouns in the SYN2015 corpus. It shows the standard behavior of Czech nouns, as well as the threshold for an unusually high proportion of each of the cases. This threshold varies notably across individual cases, e.g., 2.5% is an unusually high proportion of dative plural, whereas 24.1% is an unusually high proportion of nominative plural, and the percentage is even higher (57.4%) for nominative singular. As mentioned above, the lower threshold for all cases is zero, in other words, a gap in the paradigm (vocative case is an exception). Specific values relating to the boxplots in figure 1 (median, interquartile range, and outliers) are presented in table 1.

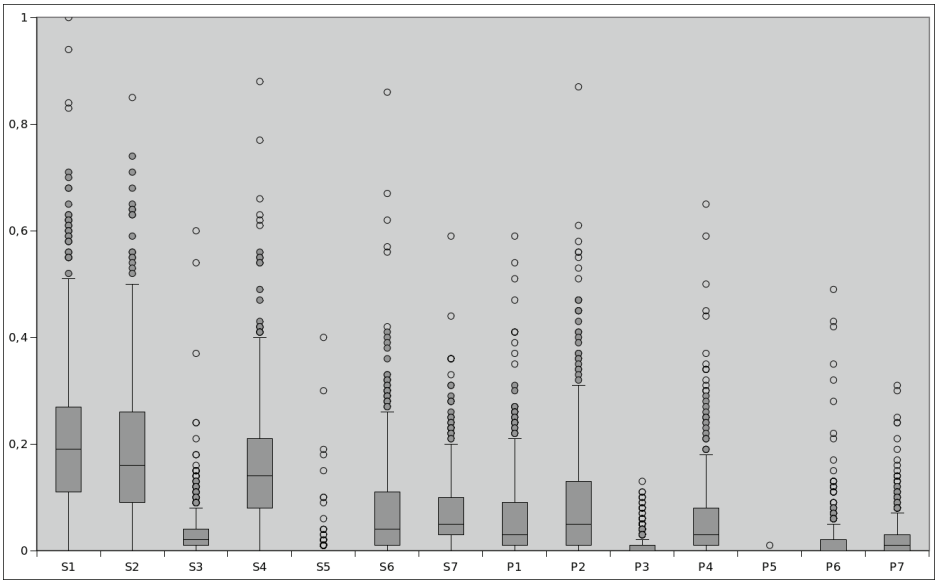


Fig. 1. Case distribution of Czech nouns in the SYN2015 corpus (nouns with a frequency of at least 100). Outliers indicate individual noun lemmas that have an unusually high percentage of the given case. Boxes, together with whiskers, represent the range of standard behavior of nouns

[5] We look at distribution in of 14 cases, i.e., 7 cases in two numbers, to capture the whole paradigm of each lemma.

	singular							plural						
	1 (nom)	2 (gen)	3 (dat)	4 (acc)	5 (voc)	6 (loc)	7 (inst)	1 (nom)	2 (gen)	3 (dat)	4 (acc)	5 (voc)	6 (loc)	7 (inst)
Unusually high	57.4	54.9	9.3	48.7	0.0	25.2	21.6	24.1	29.0	2.5	19.1	0.0	4.2	7.4
75th perc.	28.0	25.0	4.0	22.6	0.0	10.5	9.8	9.8	11.8	1.0	7.8	0.0	1.7	3.0
Median	19.3	14.8	1.8	14.5	0.0	3.9	5.5	3.5	4.2	0.2	2.8	0.0	0.4	0.9
25th perc.	12.5	7.4	0.7	7.8	0.0	0.9	2.8	0.5	0.5	0.0	0.3	0.0	0.0	0.0
Unattested	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Tab. 1. Supplementary table to Fig. 1. The value in the first line indicates the threshold for outliers or the threshold of an unusually high proportion of the given case (in %). Values between the 25th and 75th percentile form the box in the boxplot for each case. The values are calculated based on nouns that occurred with a frequency of at least 100 in the SYN2015 corpus

The overview of standard behavior of nouns within the grammatical category of case is not completely new information, it has been in shorter form presented in the books *Statistiky češtiny* [20, p. 134] and *Mluvnice současné češtiny* [16, p. 141]. Also, it is not difficult to extract this information directly from a corpus concordancer such as KonText. But this basic information is merely a gateway to grammatical profiles of all 14,000 lemmas examined, as well as to the groups of lemmas belonging to outliers (see section 4.2).

4.2 Grammatical profiles of individual lemmas

In GramatiKat, it is possible to display the grammatical profile of a particular lemma against the background of standard behavior of the whole part of speech. In figure 2, we can see the case distribution of the lemma *sekera* ‘ax’ in the form of grey dots, figure 3 shows the data for the word *uvozovka* ‘quotation mark’. In both figures, there is an evident deviation from the standard. Figure 2 shows an unusually high frequency of instrumental singular (high percentage of instrumental is characteristic of other tools as well, such as *lopata* ‘spade’, *kladivo* ‘hammer’, *nůž* ‘knife’, or *hrábě* ‘rake’). In figure 3, we can see that the lemma is overall more common in the plural, and we can observe an extremely high frequency of locative plural (*v uvozovkách* ‘in quotation marks’).

Finding words that have an unusually high percentage of a certain case is also possible. For example, under dative singular, we can find 1169 lemmas where this case accounts for at least 9.3% (the threshold for outliers, see table 1). The lemmas that occur almost exclusively in this case include *mání* ‘having’, *dostání* ‘getting’, *nepoznání* ‘not recognizing’, *zahození* ‘discarding’, or **snědek* ‘eating’. All of these lemmas are components of multiword units (mostly with the verb *být* ‘to be’ and preposition *k* ‘to’: *ne/být k mání* ‘not/to be had’, *ne/být k dostání* ‘not/to be gotten’, *být k nepoznání* ‘to be unrecognizable’, *něco k snědku* ‘something to be eaten’) and their classification as nouns is entirely formal. This is especially evident in the reconstructed

nominative singular **snědek*. Among other lemmas with unusually high but not exclusive dative singular (around 20%) are *jubileum* ‘anniversary’, *politování* ‘regret’, *zlepšení* ‘improvement’, *usmrcení* ‘killing’, and *obezřetnost* ‘prudence’.

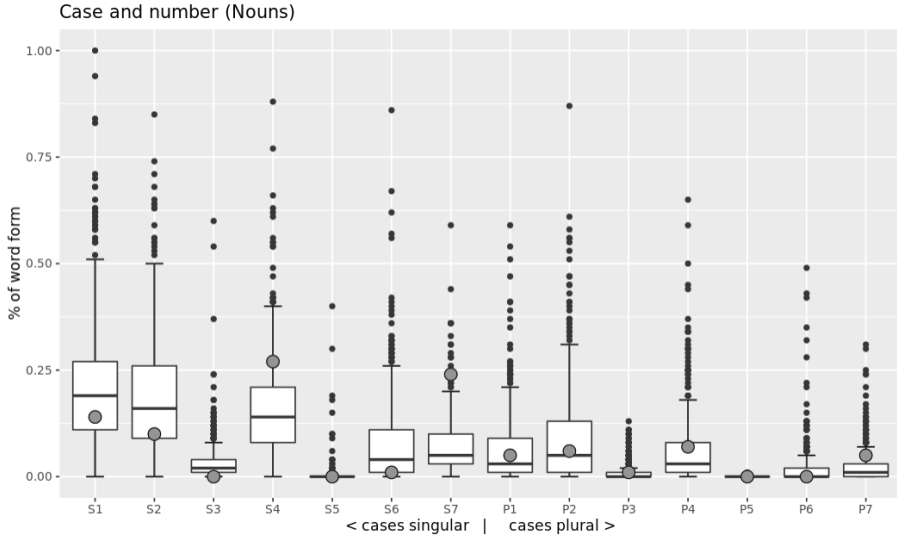


Fig. 2. The grey dots show the percentage of individual cases within the lemma *sekera* ‘ax’, the background boxplots show standard behavior as well as outliers of Czech nouns

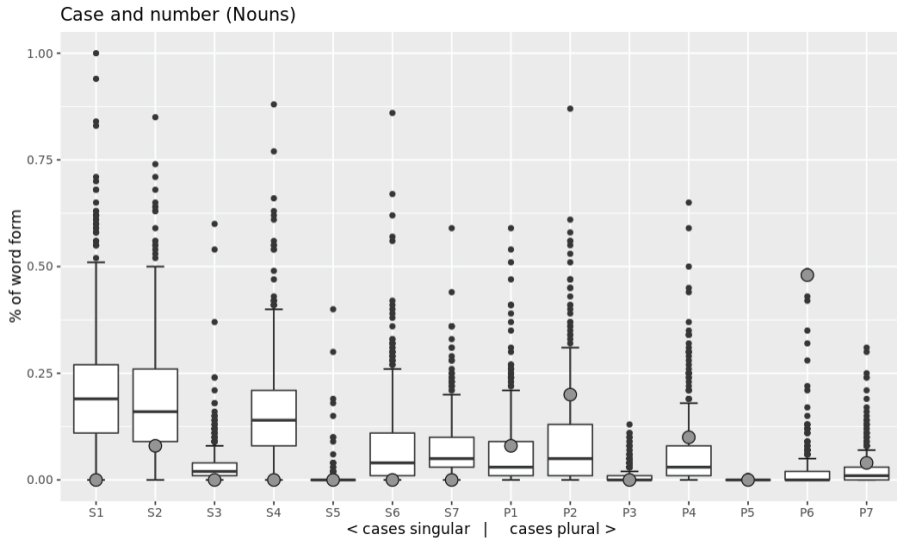


Fig. 3. The grey dots show the percentage of individual cases within the lemma *uvozovka* ‘quotation mark’, the background boxplots show standard behavior and outliers of Czech nouns

Similarly, it is possible to find lemmas with a missing form, for example, nominative plural. Almost 25% of the examined nouns, or 3400 lemmas, do not occur in nominative plural.[6] Such a comprehensive list of singularia tantum can lead to a better understanding and theoretical description of this phenomenon, especially the reasons for missing plural forms (usually semantic incompatibility, strong semantic preference, or limited collocability [21, p. 6]. Some of the lemmas with the plural form missing are *agresivita* 'aggressiveness', *bezpečí* 'safety', *komplexnost* 'complexity', *počasí* 'weather', or *potomstvo* 'offspring'.

4.3 Proportion of standard behavior nouns

The common presumption that most of the reasonably frequent nouns have a complete paradigm with no significant deviations is revealed as incorrect. On the contrary, the examination of the material available in GramatiKat shows that only about 25% of nouns with a frequency of at least 100 in the corpus can be considered standard concerning the distribution of cases[7] – all the cells of their paradigms are represented and there are no unusually frequent paradigm cells.[8] More specifically, approximately half of the lemmas examined show an unusually high frequency of at least one paradigm cell, and approximately 50% of the lemmas show at least one missing paradigm cell, with a significant overlap between the two groups.

However, this phenomenon is highly frequency-sensitive. The percentage of standard lemmas increases (up to a point, see figure 4) and decreases with their frequency in the corpus – the probability of attested dative plural, for example, is quite low in lemmas with a frequency lower than 100. And ultimately, a lemma with a frequency lower than 14 cannot be represented by all 14 cases.

In any case, non-standard behavior in nouns is not a marginal phenomenon but rather a frequent feature that should be monitored and described not only within the realm of grammar but also in lexicographical description (see section 5).

4.4 Comparison of Czech and Slovak nouns

The GramatiKat tool is ready to process material from languages other than Czech as well. The prerequisite is a sufficiently large morphologically tagged corpus. We have so far processed Croatian nouns (available upon request) and Slovak nouns. A major issue for comparing two languages, as well as for reliability

[6] Such lemmas do not occur or rarely occur in any of the plural forms.

[7] L. Janda and F. M. Tyers claim that "[o]nly a fraction of lexemes are encountered in all their paradigms in any corpus or even in the lifetime of any speaker" [18, p. 1]. The results presented here show that the situation is not as severe (perhaps the corpus size played a role). However, it is possible to agree that non-standard paradigms are not an exception.

[8] Again, the vocative was excluded.

of the results, is their dependency on corpus size, types of texts or, in the case of smaller corpora, even on the individual texts included. For Czech, we are satisfied with working with the representative and balanced corpus SYN2015. For other languages including Slovak, InterCorp data are large and diverse enough (even though not balanced). They offer the possibility to compare two (and even more) languages on the basis of the exact same texts, which we implemented in the GramatiKat tool for the language pair of Czech and Slovak. The comparison of Czech and Slovak is very reliable, the results for the two separate languages are less so (compare figure 1 with 5 and 6).

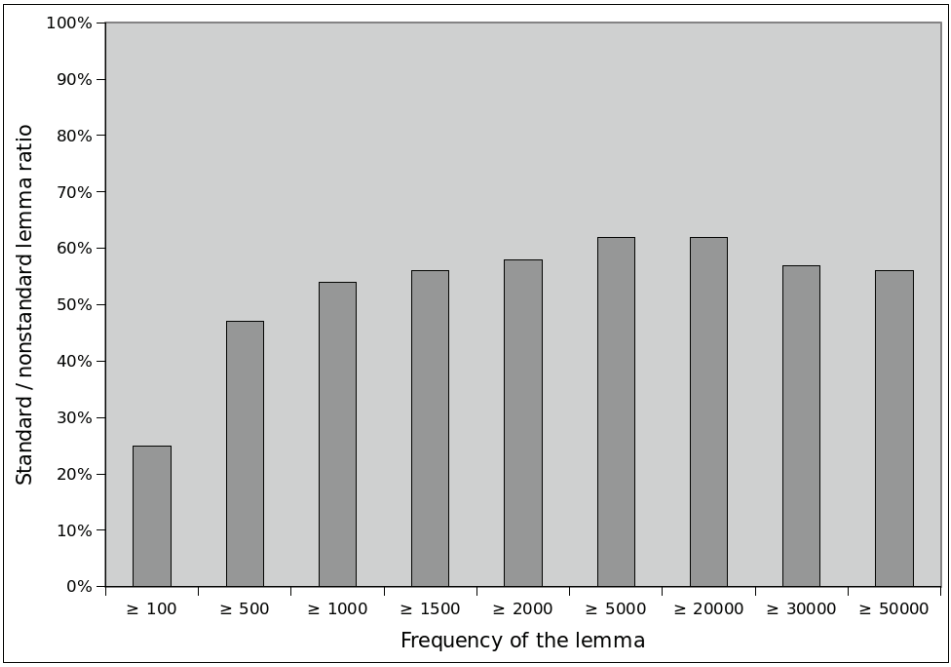


Fig. 4. Percentage of lemmas with a standard grammatical profile on different frequency levels. The figure shows that the phenomenon is frequency-dependent

A comparison of case distribution in Czech and Slovak (figure 5 for singular and figure 6 for plural, also summarized in table 2) shows that the two languages are very close in this respect. The biggest differences are between nominative singular, which is 1.1 percent more frequent in Czech, whereas accusative singular is 0.9 percent more frequent in Slovak. Whether or how these two phenomena are related to each other could only be determined through further extensive analysis.

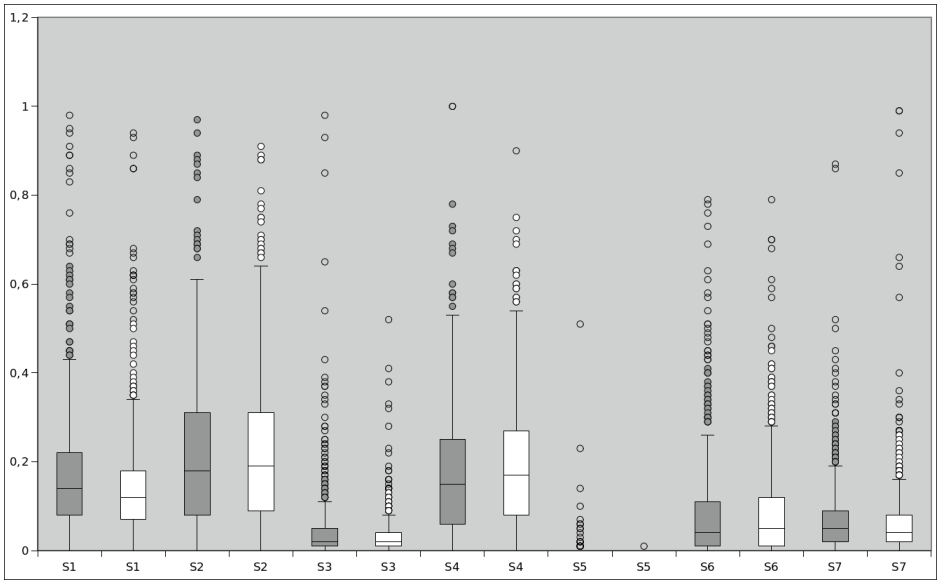


Fig. 5. Comparison of singular cases distribution in Czech (grey) and Slovak (white) nouns in InterCorp version 13, lemmas with frequency of at least 100

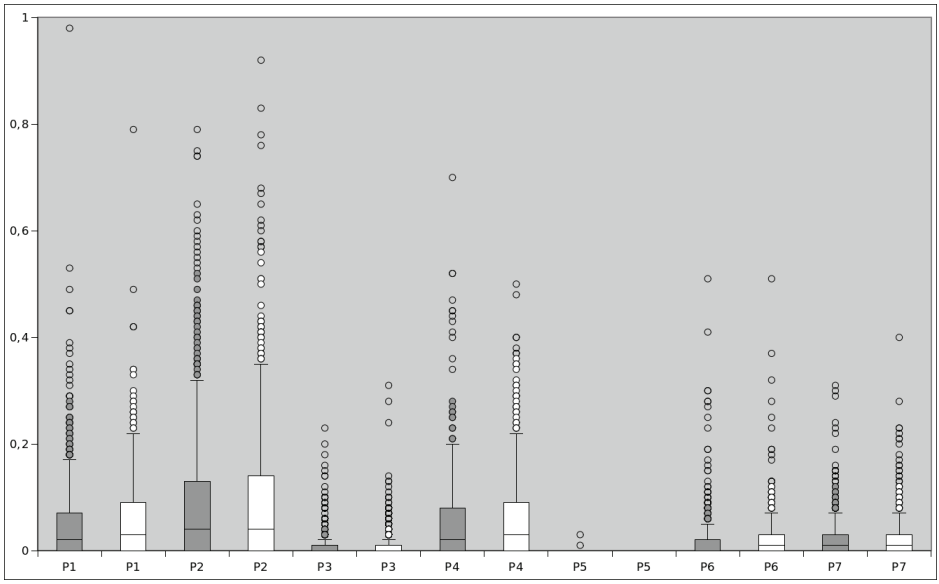


Fig. 6. Comparison of plural cases distribution in Czech (grey) and Slovak (white) nouns in InterCorp version 13, lemmas with frequency of at least 100

	singular							plural						
	1 (nom)	2 (gen)	3 (dat)	4 (acc)	5 (voc)	6 (loc)	7 (inst)	1 (nom)	2 (gen)	3 (dat)	4 (acc)	5 (voc)	6 (loc)	7 (inst)
Median CZ	13.35	17.28	2.07	15.13	0.00	3.87	4.65	2.35	3.61	0.16	2.70	0.00	0.42	0.72
Median SK	12.24	17.65	1.66	16.04	0.00	4.67	4.73	2.53	3.85	0.08	2.94	0.00	0.51	0.74
Difference	**-1.11**	**0.37**	**-0.40**	**0.91**	**0.00**	**0.80**	**0.08**	**0.18**	**0.23**	**-0.08**	**0.24**	**0.00**	**0.09**	**0.01**

Tab. 2. Supplementary table to fig. 6 and 7 showing the difference between Czech and Slovak standard noun behavior. The values (in %) are calculated based on nouns that occurred with a frequency of at least 100 in the InterCorp version 13 parallel Czech-Slovak subcorpus

5 GramatiKat IN LINGUISTIC RESEARCH – SUGGESTIONS

The GramatiKat data can be utilized in various linguistic disciplines. Instant use is possible in lexicography by detecting the lemmas with non-standard behavior (gaps in paradigm, extremely frequent forms). For example, it could be helpful to supplement the entry *brva* 'eyelash' in the *Academic Dictionary of Contemporary Czech* [22] with the information that 77% occurrences of this lemma are in instrumental singular, so the lemma is overwhelmingly often a component of the idiom *ne(po)hnout (ani) brvou* 'not to bat (even) an eyelash' (the idiom itself is listed in the dictionary, without frequency information).

Similarly, the tool can be used for educational purposes, especially in teaching Czech as a second language. Adaptation of educational practices based on case distribution is discussed by Janda and Tyers [18] who suggest that "learning may be enhanced by focusing only on the word forms most likely to be encountered" [18, p. 28). For example, we can consider teaching only the genitive and accusative singular of the lemma *večer* 'evening' (nominative, genitive and accusative represent 78% of the lemma occurences), and genitive and locative singular of the lemma *zahrada* 'garden' (64% of occurrences), especially in the earlier stages of the learning process.

The obvious direction for closer examination of the pre-processed data is morphological analysis. Determination of quantitative properties of individual grammatical categories within the individual parts of speech alone can be a valuable outcome. With information on all grammatical categories completed, we can expect re-evaluation or more accurate understanding and description of morphological phenomena (as was the case of adjective gradation mentioned above). Since the anomalies in case distribution are often caused by collocational restrictions, research should be also oriented toward multi-word units which are underrated and underrepresented in current grammatical, as well as lexicographic descriptions.

As a part of the Feast and Famine project[9], research of defectivity and anomalies in grammatical profiles of Czech nouns is currently underway. The preliminary results show that GramatiKat data is very relevant to theoretical research of language potentiality and of paradigm defectivity, as well as the underlying motives (especially semantics and collocability).

6 CONCLUSION

This article is based on the plenary session of the Slovko 2021 conference. It presents an online tool for research of grammatical categories – GramatiKat. The tool reflects the current atmosphere of open access and shared data in science and humanities, as noted in Chromý and Cvrček [1]. It provides users interested in linguistic research of grammatical categories (namely in the fields of morphology and lexicography) with a large-scale, pre-processed corpus data, as well as visualizations of grammatical categories of Czech. Also available is a comparison of Czech and Slovak nouns based on a parallel corpus of the two languages.

The study gives an overview of the grammatical category of case (in combination with number) in Czech nouns – it shows the standard behavior of Czech nouns, as well as the thresholds for non-standard case distribution. On this basis, the charts in GramatiKat also show the case distribution and anomalies within paradigms of individual lemmas. The anomalies are not a peripheral phenomenon within nouns; section 4.3 shows that a significant number of lemmas exhibit non-standard case distribution – either a paradigm gap or an unusually high frequency of a certain case.

However, the main goal of this article is not to present the tool itself (although as a co-author, I am grateful for this opportunity); my ambition is to inspire others to undertake similar projects which provide other linguists with otherwise inaccessible data and facilitate a broader and deeper examination of a specific phenomenon. I demonstrated in several examples how a tool such as GramatiKat can be versatile and can serve researchers of various linguistic fields or interests. The data is relevant to morphology, as well as lexicology and lexicography, to theoretical research of language potentiality and defectivity, and can be also used for educational purposes.

The benefits of a tool such as GramatiKat, offering pre-processed data, are numerous. The shared data follows the principles of Open Science, namely the verifiability of research and building on previous research. Most importantly, such tool gives all linguists, corpus and non-corpus, access to data that might otherwise be unattainable. The users can then undertake thorough research of a scale that is impossible for one person and can also use the tool in original and unexpected ways.

[9] Feast and Famine: Confronting Overabundance and Defectivity in Language is a project that takes place in several European universities and language institutes, including Sheffield University, the Faculty of Arts of Charles University and the Czech Language Institute (https://www.sheffield.ac.uk/feastandfamine).

ACKNOWLEDGEMENTS

This paper resulted from implementation of the Czech National Corpus project (LM2018137) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures. The research has been also in part funded by the United Kingdom's Arts and Humanities Research Council (AH/T002859/1) and by the European Regional Development Fund-Project "Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World" (No. CZ.02.1.01/0.0/0.0/16_019/0000734).

References

[1] Chromý, J., and Cvrček, V. (2021). Lingvistika jako otevřená a transparentní disciplína. Naše řeč, 104(1), page 514.

[2] Cvrček, V., and Kováříková, D. (2011). Možnosti a meze korpusové linvistiky. Naše řeč, 94(3), pages 113–133.

[3] Kováříková, D., and Kovářík, O. (2021). GramatiKat. Prague: ÚČNK FF UK. Praha 2021. Accessible at: http://www.korpus.cz/gramatikat.

[4] Cvrček, V., and Vondřička, P. (2011). SyD – Korpusový průzkum variant. Prague: FF UK. Accessible at: http://syd.korpus.cz.

[5] Cvrček, V., and Vondřička, P. (2013). Morfio. Prague: ÚČNK FF UK. Accessible at: http://morfio.korpus.cz.

[6] Cvrček, V., and Vondřička, P. (2013). KWords. Prague: ÚČNK FF UK. Accessible at: http://kwords.korpus.cz.

[7] Vavřín, M., and Rosen, A. (2015). Treq. Prague: ÚČNK FF UK. Accessible at: http://treq.korpus.cz.

[8] L. Lukešová (ed.). (2017). Pro školy – reportáž korpusových cvičení. Prague: ÚČNK FF UK. Accessible at: http://www.korpus.cz/protokoly.

[9] Machálek, T. (2019). Slovo v kostce – agregátor slovních profilů. Prague: ÚČNK FF UK. Accessible at: http://www.korpus.cz/slovo-v-kostce.

[10] Cvrček, V. (2019). Calc: Korpusová kalkulačka. Prague: ÚČNK FF UK. Accessible at: http://www.korpus.cz/calc.

[11] Křen, M., and Cvrček, V. (2019). Lists: Prohlížeč frekvenčních seznamů. Prague: ÚČNK FF UK. Accessible at: http://www.korpus.cz/lists.

[12] Vondřička, P. (2020). KorpusDB: Databáze slovních tvarů a lemmat doložených v korpusech ČNK. Verze 1.0. Prague: ÚČNK FF UK. Accessible at: http://db.korpus.cz/.

[13] Cvrček, V., Čech, R., and Kubát, M. (2020). QuitaUp – nástroj pro kvantitativní stylometrickou analýzu. Czech National Corpus and University of Ostrava. Accessible at: https://korpus.cz/quitaup/.

[14] Goláňová, H., Waclawičová, M., and Pejcha, J. (2021). Mapka: Mapová aplikace pro korpusy mluvené češtiny. Verze 1.1. Prague: ÚČNK FF UK. Accessible at: http://www.korpus.cz/mapka.

[15] Kováříková, D., and Kovářík, O. (2021). Akalex. Prague: ÚČNK FF UK. Praha 2021. Accessible at: http://www.korpus.cz/akalex.

[16] Cvrček, V. et al. (2009). Mluvnice současné češtiny I.: Jak se píše a jak se mluví. Praha: Karolinum.
[17] Křen, M. et al. (2020). SYN2020: reprezentativní korpus psané češtiny. Prague: ÚČNK FF UK. Accessible at: http://www.korpus.cz.
[18] Janda, L. A., and Tyers, F. M. (2018). Less is more: why all paradigms are defective, and why that is a good thing. Corpus linguistics and linguistic theory, 14(2). Accessible at: https://doi.org/10.1515/cllt-2018-0031.
[19] Křen, M. et al. (2015). SYN2015: reprezentativní korpus psané češtiny. Prague: ÚČNK FF UK. Accessible at: http://www.korpus.cz.
[20] Čermák, F. et al. (2009). Statistiky češtiny. Prague: NLN.
[21] Kováříková, D. et al. (2019). Lexicographer's Lacunas or How to Deal with Missing Representative Dictionary Forms on the Example of Czech. International Journal of Lexicography, 33(1), pages 90–103. Accessible at: https://doi.org/10.1093/ijl/ecz027.
[22] Akademický slovník současné češtiny (2021). Accessible at: https://slovnikcestiny.cz/uvod.php.

sciendo
DOI 10.2478/jazcas-2021-0050

THE NEW VALUE OF THE STRUCTURAL ATTRIBUTE *SECTION* IN THE SYN v8 CORPUS AND ITS POSSIBLE APPLICATION IN LINGUISTIC RESEARCH

ZUZANA LAUBEOVÁ – MICHAL ŠKRABAL
Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic

LAUBEOVÁ, Zuzana – ŠKRABAL, Michal: The new value of the structural attribute *section* in the SYN v8 corpus and its possible application in linguistic research. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 545 – 555.

Abstract: The paper introduces a new section separated from journalistic texts in Czech corpora, namely interviews. This genre is highly specific; from among the texts that can be found in newspapers and magazines, it is probably the closest to spoken language. In two case studies, we present the possible application of the interviews subcorpus in linguistic research. The first one deals with the role of paralinguistic behaviour, especially laughter in written interviews vs. spoken dialogues. The second one investigates the specifics of the demonstrative *ten* in the function of a nominal attribute, again in both written and spoken data.

Keywords: Czech spoken corpora, interviews, paralinguistic behaviour, determiner *ten*

1 INTRODUCTION

Language corpora contain many metadata which help to analyse the actual data. In general, information about written texts is divided according to the whole document, the text itself, the paragraph, and the sentence, and is generally referred to as containing structural attributes. Each of them is filled with different values. This paper aims to introduce the structural attribute of a *section,* covering the content of a newspaper or a magazine (hereinafter: NMG) split into the sections, e.g., news, sport, crime, etc. The attribute value is based on the original newspaper; therefore, it can vary from one title to another or stay empty, unfilled.

The attribute of *section* has been introduced in the corpus of written Czech SYN2015 [1]. Besides 13 original values of this attribute, another one – *rozhovory* (‘interviews’) – was added in the SYN v8 corpus [2]. This genre is highly specific; of most of the texts in NMG, it is probably the closest to spoken language – albeit admitting that interviews are edited and “smoothed” towards the easy-to-read form. Moreover, it is easily detectable too (most usually titled *Rozhovor s…* ‘Interview with’), and as such, it served us well as a starting dataset for our research. Among other things, it turned out that the information about the interviewee’s behaviour is

usually added in the parentheses, and these paralinguistic comments (e.g., *směje se* 'laughs', *krčí rameny* 'shrugs his/her shoulders', etc.) can specify the interpretation of the speaker's verbal message. We present the results and compare them with the transcribers' comments within the Czech spoken corpora in the first case study of our paper. The second case study deals with the demonstrative *ten* ('the') in the function of a nominal attribute in both written and spoken data.

2 CASE STUDY 1: PARALINGUISTIC COMMENTS IN NMG INTERVIEWS

2.1 Introduction

The first case study deals with the role of paralinguistic behaviour, especially laughter, in written interviews vs. spoken dialogues. Especially recently, the interviewee's nonverbal behaviour is sometimes captured in NMG interviews (hereinafter: NMGi) to specify and/or complete the meaning of the utterance. It is similar to the author's comments about the behaviour of characters in a script or a play.

What the comment is comprised of and what is essential to mention depends on an interviewer (or an editor). There are no strictly given rules or requirements. We can only assume that there is a common practice of processing the spoken transcript of an interview into a written form that is different in each newspaper or magazine, and this practice also includes the use of comments.

This study follows up and expands on the unpublished pilot study [3], conducted on uncategorized journalistic texts from the SYN v6 corpus [4]. Preliminary results showed that the three most frequent comments (laughter, smile, thinking) covered 95% of all comments (see Table 4). From the formal point of view, comments are mainly composed of a verb, a verb with an adverb, or, alternatively, a more complex structure (e.g., *začne se hlasitě smát* 'starts to laugh loudly'). Aiming to verify to what extent these results are valid only for the written interviews, we used the new section attribute within the SYN v8 corpus [2]. Further, the results from both written corpora are compared to the transcribers' paralinguistics comments included in the spoken corpora of Czech.

In general, our paper aims to reveal which types of comments are incorporated the written interviews and how they are structured. We also consider the overall motivation of paralinguistic comments in the texts.

2.2 Data and methodology

This study is based on three corpora of today's Czech. Within the context of the written corpora SYN v6 and v8, we focus on the journalistic texts only, as we presume a higher frequency of paralinguistic comments there. SYN v6 has no particular category for written interviews; therefore, we had to work with the whole journalistic subcorpus (4.36G), using the following CQL query:

[word="\("][word="(?i)[aábcčdďeéěfghiíjklmnňoópqrřsštťuúůvwxyýzž]{1,20}" & pos!="[XC]" & tag!=".{14}8."]
On the contrary, the NMGi subcorpus (over 2M tokens) can be delimited within the newer SYN v8 corpus, and the CQL query syntax is much more straightforward:
[word="\("] within <text section="rozhovory" />
The ORTOFON v2 corpus [5] (2.5M) represents the synchronic spoken language. We benefit from the fact that the comments are tagged as the "M" part of speech.
[pos="M"]

The results from both written corpora were manually filtered, focusing on the search for relevant results, i.e., the comments which describe the interviewee's paralinguistic behaviour.

2.3 Results

Firstly, we compare the content of comments within the SYN v6: NMG subcorpus with the SYN v8 corpus. Although there are more than 6M hits, most of them needed to be filtered out. Table 1 shows the ten most frequent types. The relevant occurrences are in bold.

	SYN v6: NMG	**ipm**
1.	*(na snímku)* 'on the photograph'	28.1
2.	*(vlevo)* 'on the left'	19.8
3.	***(smích)*** **'laughter'**	15.4
4.	*(vpravo)* 'on the right'	14.5
5.	***(směje se)*** **'is laughing'**	6.6
6.	*( )* [website removed]	5.3
7.	*(ne)* 'no/non-'	4.9
8.	***(úsměv)*** **'smile'**	4.4
9.	*(uprostřed)* 'in the middle'	4.0
10.	*(ANO)* [abbreviation of Czech political party]	3.9

Tab. 1. Top 10 most frequent chunks in parentheses in the SYN v6: NMG subcorpus

Table 1 illustrates what kind of information is mainly captured within parentheses as comments in NMG. These results are similar in both written corpora.

Besides paralinguistic behaviour, it can be a caption of a photograph or a picture (nr. 1, 2, 4, 9 in Table 1), a quotation of a website (nr. 6), or one's affiliation to a political party (nr. 10). Also, the negation particle *ne* (nr. 7) was found in texts quite often, e.g., *Vláda (ne)schválila daňovou reformu* 'The government has (not) approved the tax reform'.

Filtered results show the prevalence of laughter or smiles within both corpora (see also Table 2 below). Looking closer, we could identify the varied modes/phases of laughing and smile and/or different length of their duration. The third most frequent behaviour is the process of thinking, also of various lengths or efforts. Other types of comments describe pauses, gestures, facial expressions, or sighs – in general, the nonverbal physical or physiological behaviour. There are also the interviewer's comments on the interviewee's speech (e.g., *hledá v mobilu* 'is searching on the mobile' *skáče/skočí do řeči* 'is cutting in', *nesouhlasně vrtí hlavou* 'shakes his/her head in disapproval') and noises from outside (a phone ringing), other people's reactions (*jeho žena přikyvuje* 'his wife nods'). The mental state of the speaker is described by adverbs (e.g., *smutně* 'sadly', *pobaveně* 'amusedly', *zklamaně* 'disappointedly'), which may be added to the verbal comment, too.

rank	SYN v6: NMG	ipm	SYN v8: NMGi	ipm
1.	*(smích)* 'laughter'	15.4	*(smích)* 'laughter'	145.3
2.	*(směje se)* 'is laughing'	6.6	*(směje se)* 'is laughing'	50.8
3.	*(úsměv)* 'smile'	4.4	*(úsměv)* 'smile'	30.3
4.	*(usmívá se)* 'is smiling'	2.1	*(usmívá se)* 'is smiling'	9.8
5.	*(usměje se)* 'smiles'	1.3	*(skáče do řeči)* 'cuts in'	2.8
6.	*(rozesměje se)* 'laughs'	0.3	*(ukazuje něco/na něco)* 'points at sb/sth'	1.4
7.	*(přemýšlí)* 'is thinking'	0.2	*(usměje se)* 'smiles'	0.9
8.	*(pousměje se)* 'half-smiles'	0.2	*(se smíchem)* 'with laughter'	0.9
9.	*(zamyslí se)* 'thinks'	0.1	*(vehementně předvádí)* 'demonstrates vehemently'	0.5
10.	*(důrazně)* 'strongly'	0.1	*(nesouhlasně vrtí hlavou)* 'shakes his/her head in disapproval'	0.5
11.	*(odmlčí se)* 'falls silent'	0.1	*(přemýšlí)* 'is thinking'	0.5

rank	SYN v6: NMG	ipm	SYN v8: NMGi	ipm
12.	*(s úsměvem)* ‘with a smile’	<0.1	*(rozesměje se)* ‘laughs’	0.5
13.	*(zasměje se)* ‘laughs’	<0.1	*(hledá v mobilu)* ‘is searching on the mobile’	0.5
14.	*(skočí do řeči)* ‘cuts in’	<0.1	*(zamýšlí se)* ‘think about sth’	0.5
15.	*(kroutí hlavou)* ‘shakes one’s head’	<0.1	*(s úsměvem)* ‘with a smile’	0.5
16.	*(chvíli přemýšlí)* ‘is thinking for a while’	<0.1	*(pokyvuje hlavou)* ‘nods’	0.5
17.	*(dlouho přemýšlí)* ‘is thinking for a long time’	<0.1	*(dlouze přemýšlí)* ‘is thinking for a long time’	0.5
18.	*(se smíchem)* ‘with laughter’	<0.1	*(smutně se usmívá)* ‘smiles sadly’	0.5
19.	*(povzdechne si)* ‘sighs’	<0.1	*(chvilku přemýšlí)* ‘is thinking for a while’	0.5
20.	*(skáče do řeči)* ‘is cutting in’	<0.1	*(úsměv na tváři mluvčího)* ‘smile on the speaker’s face’	0.5

Tab. 2. Comparison of the top 20 paralinguistic comments in the SYN v6: NMG and SYN v8: NMGi subcorpora

The results from the written corpora (Table 2) partially correspond with the comments of paralinguistic comments in the ORTOFON v2 corpus (Table 3). These comments are added during the process of transcription and focus on the sounds that could influence spontaneous dialogue. There are not only paralinguistic, nonverbal expressions of a speaker, such as breathing in and out, but also sounds accompanying the speech, e.g., noise from the street, knocking on the door, clearing the throat, etc. Table 3 shows that laughter is the third most frequent comment and a faint smile the fifth one. This type of comment is the only one that is exactly the same as in the written data.

rank	ORTOFON v2	ipm	rank	ORTOFON v2	ipm
1.	*(nadechnutí)* ‘breathing in’	8,143	11.	*(hlasitý hovor v pozadí)* ‘loud talking in the background’	580
2.	*(rušivý zvuk)* ‘disruptive sound’	5,366	12.	*(citoslovce)* ‘interjection’	541
3.	*(smích)* ‘laughter’	4,483	13.	*(smích více mluvčích najednou)* ‘collective laughter of multiple speakers’	376
4.	*(hluk v pozadí)* ‘noise in background’	2,682	14.	*(povzdech)* ‘sigh’	362

rank	ORTOFON v2	ipm	rank	ORTOFON v2	ipm
5.	*(pousmání)* ‘faint smile’	2,531	15.	*(zvuk z rádia)* ‘sound from a radio’	339
6.	*(mlasknutí)* ‘lip smacking’	1,583	16.	*(ruch z ulice)* ‘noise from the street’	318
7.	*(cinkání nádobí)* ‘clinking dishes’	1,553	17.	*(zvuky při jídle)* ‘sounds during eating’	308
8.	*(odkašlání)* ‘clearing the throat’	690	18.	*(polknutí)* ‘swallow’	294
9.	*(vydechnutí)* ‘breathing out’	646	19.	*(mluví ke zvířeti)* ‘talks to animal’	272
10.	*(klepání)* ‘knocking’	592	20.	*(zvuk z televize)* ‘sound from TV’	266

Tab. 3. Top 20 most frequent paralinguistic comments in the ORTOFON v2 corpus

2.4 Summary

This study focused on the paralinguistic comments in written journalistic texts. The most frequent comments include laughter, smile, and thinking (Table 4).

	SYN v6: NMG	SYN v8: NMGi
laughter	72%	80%
Smile	21%	17%
thinking	2%	1%
Other	5%	2%

Tab. 4. The types of comment according to their meaning within written corpora

The frequency of laughter and smile indicates that these comments are considered essential for readers to understand the tone of the interview properly. We assume that this is a primary motivation for their incorporation into texts despite a somewhat foreign nature in the stream of speech and potential difficulties with their conversion into words.[1] The presence of paralinguistic comments in NMGi is, at least from the perspective of frequency, a typical feature for this register (similar to scripts).

[1] The laughter or smile comments are sometimes replaced with emoticons, e.g., – *Která historka vás v poslední době pobavila? – Ta, že jste si za mnou přišel pro rozhovor. :)* ‘– Which story has amused you lately? – That you came to me for an interview. :)’.

3 CASE STUDY 2: THE DETERMINER *TEN* IN NMG INTERVIEWS

3.1 Introduction

The second case study concerns the specific role of the demonstrative *ten*[2] as a determiner. Although Czech belongs to languages without a definite article, its function is repeatedly attributed to the pronoun *ten*, and there are recurring hypotheses about the gradual emergence of the category of definiteness in, predominantly spoken, Czech (as early as in 1917: [6], most recently [7]). We want to verify this hypothesis – purely quantitatively at the moment – on the spoken corpora of Czech, including their specific segment of NMGi. We limit ourselves to a quantitative analysis only, which should then be supplemented by a deeper qualitative analysis, e.g., in a similar way as described in [7].

3.2 Data and methodology

To get an insight into behaviour of the pronoun *ten* as a means of deixis, we used the following CQL query:

```
(1:[lemma="ten" & tag="..M.*"] [pos="[AP]"]{0,3} 2:[tag="N.M.*"]) |
(1:[lemma="ten" & tag="..I.*"] [pos="[AP]"]{0,3} 2:[tag="N.I.*"]) |
(1:[lemma="ten" & tag="..F.*"] [pos="[AP]"]{0,3} 2:[tag="N.F.*"]) |
(1:[lemma="ten" & tag="..N.*"] [pos="[AP]"]{0,3} 2:[tag="N.N.*"]) & 1.case=2.case
```

We are looking for a combination of the pronoun *ten* and a noun (with up to 3 potentially inserted tokens). At the same time, both expressions must match in the case (condition 1.case = 2.case) and gender (which has no attribute in the corpora, thus it is necessary to specify via a tag successively all four genders: masculine (in) animate M/I, feminine F, neuter N). We are aware that unwanted hits remain in our dataset (besides cases of actual deixis, these are, e.g., multiword expressions such as *v tom případě* 'in that case', *tou dobou* 'at the time', etc.). Still, we intentionally do not want to advance to their manual filtering as we are only interested in obtaining and comparing the relative frequency (ipm) of this phenomenon in different corpora.

3.3 Partial results

The results summarized in Table 5 correspond to our initial hypothesis: the structure is most abundantly represented in spoken data, and within them, most often in ORATOR, the corpus of monologues about which speakers are informed in advance and for which they can prepare [8]. The speaker's effort to clearly identify the noun in question and/or to emphasize it duly in the structure of their lecture can explain the higher occurrence of the structure. We cannot exclude a double deixis –

[2] The lemma *ten* (the singular nominative form for the masculine gender) also includes the forms of the feminine and neuter gender *ta* and *to*, as well as plural forms *ti*, *ty*, and *ta*. In all the corpora we use in this chapter, the lemmatization is unified, including all these forms into the given lemma.

a verbal and physical one, i.e., real pointing to the subject using a pointer or a finger (in a presentation, on a blackboard, etc.), parallel with the utterance of the demonstrative noun. Another possible explanation includes the speaker's attempt to be informal or gain more time to word an idea by adding redundant expressions. In written corpora, this structure most often appears in fiction, while there is a significant difference between the subcorpus of interviews and NMG in general (see lines 4 and 3).

(sub)corpus	size in tokens	hits	frequency (ipm)
SYN2020: FIC	33.3M	188,013	1,543
SYN2020: NFC	33.3M	46,811	384
SYN2020: NMG	33.3M	53,745	441
SYN v8: NMGi	2.2M	6,069	2,827
ORTOFON v2	2.1M	20,794	8,121
ORAL v1	5.4M	53,500	8,410
ORATOR v2	1.2M	17,296	11,263

Tab. 5. The frequency of *ten* structures in the selected Czech corpora

Let us start with spoken language represented by the ORTOFON v2 corpus. The speaker in example (1) has a prominently higher ipm of the *ten* structure compared to ipm for the whole corpus (13,115 vs. 8,121), and it is obviously a salient feature of his idiolect.

(1) 15T008N, Ignác V. [talks about his visit to Jerusalem and comments on the behavior of Orthodox Jews at the Wailing Wall]
*takže **ty** bezpečnostní **ty** tam jsou jako hodně no tak jsme prošli .. a pak jsme byli fakt u **tý** Zdi nářků a tam to je úplně teda speciální tam .. fakt choděj jenom **ti** ortodoxní Židi takový **ty** dlouhý kabáty .. fakt choděj jenom **ti** ortodoxní Židi takový **ty** dlouhý kabáty .. a prostě u **té** Zdi nářků všichni tak jako se kolébají a říkají **tu** modlitbu*

On the contrary, the next speaker's utterance is laconic in terms of the frequency of the pronoun *ten* (ipm 2,575):

(2) 20A011N, Dušan M.[3] [depicts a difficult traffic situation at an intersection while driving a car]

[3] There is an incorrect piece of information in the database, this is in fact a female speaker. Potential yet unrealized occurrences of the pronoun are denoted by **(0)**, or **(0?)** in questionable cases, respectively.

tak jsem jela a říkám Honzíku tak . dobrý .. to snad dám .. snad to nikde tatínkovi nepoškodím .. bude to dobrý zavezu tě do **(0)** *školy . maminka tě odveze nakoupí přijede dom .. no nicméně jsem přijela .. do Ole* @ do* **(0?)** *Olešnice na* **(0?)** *křižovatku .. @ viděla jsem že jede . velkej traktor tak říkám .. tak . ho pustím udělám dobrej skutek pustím ho .. pustila jsem* **(0)** *traktor .. rozjela jsem se . a* **(0)** *auto mně chcíplo . uprostřed* **(0)** *křižovatky … za mnou auto vedle mě auto vedle mě dělníci všichni se mi @ mně smáli já jsem .. červenala začala jsem .. panikařit t* .. nezačala jsem panikařit začala jsem nadávat proč mi* ***to*** *auto půjčuje ..*

The following sample comes from the ORATOR v2 corpus, which shows the highest ipm (11,263) within the spoken corpora.

(3) 18X058F, Pavelka S. (ipm 17,670) [comments on a graph showing a decreasing amount of exercise for current children compared to previous years]
zkrátka ukazuje se asi to že tady někde okolo ***toho*** *roku devadesát šest .. už to množství* ***těch*** *realizovanejch pohybovejch aktivit .. kleslo pod* ***tu*** *biologickou potřebu .. a teď už jsme v období . kdy jenom se zhoršujeme .. otázka je jak se z toho dostat dál a jak tomu .. asi .. u jednoho 3 .. co jsou doporučení vokolo* ***toho*** *pohybu ..*

Naturally, we do not find spoken language only in spoken corpora. With some retreat from authenticity, we can find it also in written corpora, chiefly in fictional speech (as opposed to narrative). No matter how successful a writer is in pursuit of representing genuine speech, it is easily detectable in the corpus, be it by quotation marks or similar means. The situation in NMGi is different. Although we cannot be entirely sure (if we did not directly hear the speech from the authentic recording) to what extent it was corrected, paraphrased, re-stylized, etc. by an interviewer, an editor, or a proofreader, a general idea of the degree of authenticity of the quoted statement can often be made. As in the following example from NMGi (along with examples (1) and (3)), it shows occurrences of the pronoun *ten* not really corresponding to its primary (i.e., deictic) function:

(4) Sedmička, č. 38/2018 [a singer presents his new album]
A ještě musím k ***té*** *desce říct jednu věc . Je důležité , že to zpívám v češtině . Teď je trend zpívat v angličtině a* ***ty*** *věci v ní samozřejmě byly napsány , jsou to zahraniční věci , ale moje specialita je zpívat v češtině . Některé věci se tak k lidem dostanou snáz . Líp to pochopí . Navíc* ***ta*** *práce s češtinou je o hodně těžší než to zpívat anglicky . V angličtině to jde samo . Zazpíváte větu a jste světový zpěvák . Ale u* ***té*** *češtiny musíte prokázat interpretační zralost a schopnost .*

3.4 Summary

Our second topic is not a full-fledged case study, but rather a quantitative probe in regards to available corpus data. Primarily, we aimed to indicate the connection

between the use of the demonstrative *ten* and specific registers and to show a descending tendency from genuinely spoken data, where *ten* is a strong indicator of spontaneity and/or emotionality, to written data. Nevertheless, even in written registers, e.g., in NMGi, they may play a prominent role: their increased frequency contributes to possible grammaticalization of the pronoun *ten* into a definite article. Understandably, deeper qualitative analysis, preceded by manual filtration of objectionable results, is necessary.

4 CONCLUSION AND FUTURE WORK

The specificity of the NMGi register is obvious, and we must always be aware that it is an inauthentic, further modified linguistic imprint of a speaker. Therefore, it should not be confused with genuine data in spoken corpora. The extent of the interventions by an interviewer, an editor, a proofreader, or other members of editorial staff can only be speculated, and we do not know of any study that would address this issue, at least for Czech. It would certainly be illuminating to have authentic recordings available and then compare them with the final form of interviews. What do the editors consider undesirable in the interviewee's speech (in terms of content and/or language), what the reader "stand", etc.? These are only a few of the many research questions waiting to be answered.

Our paper aimed to demonstrate the usefulness of extending the list of the section attribute in the Czech NMG subcorpora by interviews, truly a specific register worthy of linguists' attention. In the first case study, we examined how the diverse types of respondents' paralinguistic behaviour are recorded in interviews and with what frequency. In the second case study, we focused on the frequency of the pronoun *ten*, which is significantly more salient in spoken utterances than in written texts. That supports the hypothesis of its gradual grammaticalization, i.e., transformation into a definite article in spontaneous spoken Czech.

The potential that such a handily defined NMG section has is far from exhausted. To look for other phenomena (e.g., contact particles or various n-grams such as *já si myslím že* 'I think that', *je/není to o* 'it is (not) about', *na druhou stranu* 'on the other side', etc.) in the NMGi subcorpus and compare them with different registers is equally tempting.

ACKNOWLEDGEMENTS

This paper resulted from the implementation of the Czech National Corpus project (LM2018137) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

References

[1] Křen, M. et al. (2015). SYN2015: reprezentativní korpus psané češtiny. Praha: Ústav Českého národního korpusu FF UK.

[2] Křen, M. et al. (2019). Korpus SYN, verze 8 z 12. 12. 2019. Praha: Ústav Českého národního korpusu FF UK. Accessible at: http://www.korpus.cz.

[3] Komrsková, Z., and Škrabal, M. (2018). The role of paralinguistic behaviour, especially laughter in written interview vs. spoken dialogue. A corpus-based study. Poster at the Second International Conference on Sociolinguistics (ICS.2), 6–8 September 2018, Budapest.

[4] Křen, M. et al. (2017). Korpus SYN, verze 6 z 18. 12. 2017. Praha: Ústav Českého národního korpusu. Accessible at: http://www.korpus.cz.

[5] Kopřivová, M. et al. (2020): ORTOFON v2: Korpus neformální mluvené češtiny s víceúrovňovým přepisem. Praha: Ústav Českého národního korpusu FF UK. Accessible at: http://www.korpus.cz.

[6] Zubatý, J. (1917). Ten. Naše řeč, 1(10), pages 289–294.

[7] Dvořák, J. (2020). The emerging definite article ten in (informal spoken) Czech: a further analysis in terms of semantic and pragmatic definiteness. Naše řeč, 103(4), pages 297–319.

[8] Kopřivová, M. et al. (2020): ORATOR v2: Korpus monologů. Praha: Ústav Českého národního korpusu FF UK. Accessible at: https://www.korpus.cz.

sciendo
DOI 10.2478/jazcas-2021-0051

AN HMM-BASED PoS TAGGER FOR OLD CHURCH SLAVONIC

OLGA LYASHEVSKAYA[1,2] – ILIA AFANASEV[3]
[1] National Research University "Higher School of Economics", Moscow, Russia
[2] Vinogradov Institute of the Russian Language, Russian Academy of Sciences, Moscow, Russia
[3] Federal State Budgetary Educational Institution of Higher Education "Saint Petersburg State University", Saint Petersburg, Russia

LYASHEVSKAYA, Olga – AFANASEV, Ilia: An HMM-based PoS tagger for Old Church Slavonic. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 556 – 567.

Abstract: We present a hybrid HMM-based PoS tagger for Old Church Slavonic. The training corpus is a portion of one text, Codex Marianus (40k) annotated with the Universal Dependencies UPOS tags in the UD-PROIEL treebank. We perform a number of experiments in within-domain and out-of-domain settings, in which the remaining part of Codex Marianus serves as a within-domain test set, and Kiev Folia is used as an out-of-domain test set. Analysing by-PoS-class precision and sensitivity in each run, we combine a simple context-free n-gram-based approach and Hidden Markov method (HMM), and added linguistic rules for specific cases such as punctuation and digits. While the model achieves a rather non-impressive accuracy of 81% in in-domain settings, we observe an accuracy of 51% in out-of-domain evaluation, which is comparable to the results of large neural architectures based on pre-trained contextual embeddings.

Keywords: HMM tagger, Old Church Slavonic, PoS tagging, hybrid models, Universal Dependencies

1 INTRODUCTION

Part-of-speech (PoS) tagging has been around for quite a long time as one of the tasks of natural language processing (NLP). Generally, the task is defined as assigning a PoS label to the token, taking into consideration lexical and contextual information. Sometimes, the tags correspond not only to the PoS categories stricto sensu, but also to the morphological features of the token [1], however, we will adhere to the former task definition. The main challenge of PoS tagging is resolving the ambiguity [2]: when considering the token sequence $w_1 \ldots w_n$, one should ideally assign one, and only one tag from a tag set $t_1 \ldots t_n$ to each token [3].

The methods of PoS tagging for different languages have achieved a certain level of sophistication. Nevertheless, the tagging of less-resourced languages and languages with considerable dialectal and local variation leaves room for experiments. Old Church Slavonic (OCS) is a language preserved in a limited

number of manuscripts, mostly ecclesiastical texts copied in monasteries in Croatia, Bulgaria, and Macedonia, that display a mixture of dialectal features. Due to diversity of language material, there is still an open discussion whether all texts belong to the same language. We were inspired by the idea of building a linguistically informed approach to PoS tagging, the results of which would remain stable on a heterogeneous set of texts. The results of computational experiments in this case are to be the subject of interpretation in linguistic terms. Which parts of speech are the hardest ones to tag? What are the particular reasons for this? Why does one method achieve higher efficiency, what makes this particular language work in terms of distribution of tokens? How linguistics may help in selecting more efficient methods of tagging? When one is able to answer these questions, one achieves another aim, the linguistic interpretability of the model.

The main method of PoS tagging presented in the article is the Hidden Markov Model (HMM), enhanced with the Viterbi algorithm ([4], [5]). The task of defining a PoS tag in HMM is reduced to the process of finding the most likely latter HMM state, while taking into consideration all the previous HMM states for all the previous observations [6]. A probable enhancement for the HMM model is a model that defines PoS if it finds one of the two most frequent n-grams that are characteristic for this particular PoS in this particular language. If the PoS is not present in the original training dataset, one might consider the application of some linguistically formed rules, specific for the corpus, used in the testing phase. This article describes different methods of adding an n-gram-based tagger to the model, as well as rules developed for some features that are not present in the training corpus, such as occurrence of fragmentary tokens, punctuation marks, and digits. All experiments are based on the Universal Dependencies (UD) UPOS tagset [7].

2 RELATED WORK

The history of PoS tagging starts with linguistic rule-based systems. This led to years of work of linguists who developed rules for a particular language. With development of technology, scientists started paying attention to a group of statistical methods ([1], [8], [9]). The following step was machine learning methods adapted to the task ([1], [5], [10], [11], [12]). Finally, recent years witnessed the appearance of taggers based on recurrent neural networks [13]. All these methods can be hybridized to form more efficient models ([11], [14], [15]). The most common method by now, though, has been HMM [3], enhanced in different ways, such as the Viterbi algorithm [6], maximum entropy method [3], or transfer learning [15]. Some methods were designed especially for tagging of extremely low-sized datasets, however, the models developed used embeddings pre-trained on bigger datasets that are impossible to get for Old Church Slavonic [16].

PoS tagging of Slavic languages witnessed a specific boom during the former two decades, see, for example, models for Czech ([17], [18]) and Russian ([19], [20], [21]). The recent years, however, sparkled the interest in the older periods of Slavic language history, provoking the appearance of taggers for Slavic languages of earlier periods, including those designed specifically for Old Church Slavonic ([22], [23]).

There are a number of corpus resources that include OCS texts. However, most of them either do not have PoS annotations (as TITUS [24]), or provide restricted access to machine-readable texts. The Manuscript project [25] has a tag set different from the UD tag set. One of the datasets described in the article is based on the UD [26] version of Codex Marianus [27].

Note that various multi-language taggers were designed to work specifically with the UD annotations (e.g., [28], [29]), since the UD repository includes material of typologically different languages annotated under similar schemas. These models have achieved significant success, with approximately a 95% accuracy score when using contextual multilingual embeddings. However, they tend to work really well on homogeneous collections of texts, large collections, and require a lot of space and resources, which make them often environmentally burdening "heavy industrial divisions". Given the fact that OCS is everything but a homogenous and large set of texts [30], they may not perform well enough on it.

3 METHOD AND DATA

For training purposes we use the full text of Codex Marianus, a version tagged for the PROIEL project [27] and then adapted and made available to UD [26]. It consists of ca. 50K tokens split into train, dev, and test parts. The train and dev parts were joined to form a training dataset, the test part was used to test the efficiency of the learning process.

As a baseline model, the n-gram counter was used. This model observed the most frequent n-grams for each PoS on the training dataset, and created a dictionary. We experimented with some enhancements, such as TF-IDF metrics and preliminary decapitalization of tokens.

Another series of training and testing was performed with the HMM model enhanced with the Viterbi algorithm [5]. After a few launches, constantly increasing the amount of data to be fed into the model, accuracy score, the metrics used for measuring the overall model performance, achieved a stable final value of 81%.

RNN-based taggers and basic regression methods were initially considered to be used as well, however, they required either significantly bigger data, or embeddings pre-trained on, once again, significantly bigger data. These models also have a tendency to overfit. This is a crucial fallacy, because the training

dataset consists only of one text, Codex Marianus, and other OCS texts may vary greatly from it. The possibility for the model to adapt to new data is of high importance, and this is why these methods were not implemented in the model.

The next stage included different methods of hybridization between HMM and n-gram models, including the regression model, training it to pick from the results of the two models. The regression model used the extra trees method, which previously proved to be efficient during different ML tracks [31]. Of all the different combinations, the hybrid of HMM and 3-gram model, with decapitalization of token in both training and testing phases, and prioritizing of adverb category assignment to HMM, proved to be the most efficient.

The final stage included out-of-domain testing. The text used was Kiev Folia, considered to be the indelible part of the OCS canon, despite having some very specific linguistic features [32].

Kiev Folia has not been tagged previously as a whole text, despite some recent attempts [22]. The original text was taken from the TITUS collection [24] and preprocessed [33]. It contained punctuation marks and digits that were intentionally deleted from ([26], [27]). For recognition of these PoS, the rule-based part of the system was implemented. The model itself is available as an open-source software [34].

4 EXPERIMENTS AND RESULTS

Four series of experiments were conducted. In the first one, two baselines were defined, using the TreeTagger [35] model trained on Bulgarian, the closest relative of OCS. The second one included a series of experiments with n-gram models. The third one presented different attempts at hybridization of HMM and n-gram models. Each of the experiments in these phases was conducted on the Codex Marianus dataset. The fourth series included testing the best model and raw HMM model on the Kiev Folia dataset.

Table 1 shows the mapping between the UD [7] and TreeTagger-Bulgarian [36] PoS tags. The Bulgarian parameters were used, since there are no trained OCS parameters for TreeTagger, and its source code, essential for the training process, is closed. The by-tag performance of the baseline model is presented in table 2. These results are the bare minimum that any model trained for OCS should beat. The baseline results are acquired via the applying of a loosely similar model, trained on a loosely similar language, mainly connected with OCS genetically, and not typologically. We also applied Russian and Slovak training parameters, however, accuracy with these achieved only 26% and 1% respectively, due to crucial differences in the tagsets.

UD	Explication	TreeTagger
ADJ	adjective	A, Mo, Md, My, H
ADV	adverb	D
INTJ	interjection	I
NOUN	noun	Nc
PROPN	proper noun	Np
VERB	verb	Vn, Vp
ADP	adposition (preposition)	R
AUX	auxiliary	Vx, Vy, Vi
CCONJ	coordinate conjunction	Cc, Cr, Cp
DET	determiner (adj. pronoun)	Ps
NUM	numeral	Mc
PART	particle	T
PRON	pronoun	Pp, Pd, Pr, Pc, Pi, Pf, Pn
SCONJ	subordinate conjunction	Cs
X	non-word	(if not tagged)

Tab. 1. Mapping UD tags onto the TreeTagger-Bulgarian

The aim of the next experiment series was to train a model that chooses the most frequent n-grams for each PoS (in the case of OCS, the two most frequent n-grams for each PoS are most helpful). During the prediction phase, the model tries to find each of these n-grams in the token, and, if not found, assigns verb, as it is the most common tag in both the training and test dataset (8356 and 2281 tokens respectively, 10637 overall). The results for different n (2, 3, 4) are presented in table 2. The 3-gram model proved to be the most efficient, though it barely outperformed the baseline, and needed further enhancements.

PoS	TreeTagger	2-grams	3-grams	4-grams
VERB	78.65	43.93	96.02	**98.88**
AUX	1.85	-	77.85	**96.04**
ADV	2.41	**70.16**	67.08	67.08
NOUN	**30.49**	16.4	-	-
PRON	4.34	66.09	83.92	**87.83**
CCONJ	36.92	**45.91**	20.66	20.66
ADP	25.97	**53.35**	49.48	49.59
ADJ	25.67	51.64	**98.8**	63.49
INTJ	8.47	**19.38**	18.12	18.12
SCONJ	2.17	31.66	**51.55**	48.97
DET	3.73	18.68	**43.9**	-
PROPN	-	**5.26**	-	-
NUM	-	-	-	-
X	-	19.66	**20.82**	20.26
Total	31.54	30.25	**32.19**	31.43

Tab. 2. Accuracy score for PoS tagging with TreeTagger-bg, 2-gram, 3-gram, and 4-gram models. The best results in comparison, here and then, are given in **bold**

N-grams were coming mostly from the first n characters of tokens of the particular PoS. The average distance from the first and the last character of the token to the first character of n-gram is provided in table 3.

PoS	Distance from the beginning	Distance from the end
VERB	0	8.1
ADJ	0.29	6.91
X	0	6
Average	0.07	7.79

Tab. 3. Average n-gram distribution for selected PoS

Then, the possibility of some enhancements to the 3-gram model was considered. For instance, we implemented a rule-based system that normalizes words that are abbreviated and covered by *titlo* (a tilde-like character) in the original texts. However, just deleting *titlo* proved to be more useful, since contractions such as *ic* (which is for the most frequent proper noun in the corpus, being *icоусъ* 'Jesus'), are more recognizable by the model.

Slight improvements of the n-gram model quality were made with decapitalization of all the tokens in the dataset. The special symbol '#' was added to the start and the end of each token, which enhanced efficiency by more than 7 per cent overall.

After that, further attempts were made, using the length of token as a criterion, counting the digraph *oy* [u] as a single vowel and ignoring repeating symbols, and introducing the TF-IDF weighting for n-grams count. It seems that token length was too ambiguous, digraphs and repeating symbols were too rare to actually influence the *status quo* for two of the most frequent n-grams, and the TF-IDF weighting did not actually make a difference for the n-grams that often. The results of the experiments are provided in the table 4.

PoS	3-gram	3 + D	3 + DB	3 + DBT	3 + DBL	3 + DBR
VERB	96.02	**96.13**	54.96	54.96	51.84	54.96
AUX	77.85	77.85	**100**	96.3	**100**	**100**
ADV	67.08	**67.39**	61.86	61.86	16.97	61.86
NOUN	-	-	**37.99**	**37.99**	36.79	**37.99**
PRON	83.92	83.92	**89.22**	**89.22**	-	**89.22**
CCONJ	20.66	21.09	**85.16**	**85.16**	46.95	**85.16**
ADP	49.48	**49.49**	40.49	40.49	-	40.49
ADJ	98.8	**98.8**	86.02	86.02	88.51	86.02
INTJ	18.12	18.57	**20.16**	**20.16**	-	**20.16**
SCONJ	51.55	51.55	78.13	78.13	**80**	78.13
DET	**43.9**	**43.9**	18.75	18.75	26.55	18.75
PROPN	-	-	93.81	93.81	**94.64**	93.81

PoS	3-gram	3 + D	3 + DB	3 + DBT	3 + DBL	3 + DBR
NUM	-	-	50	**56.9**	45	50
X	20.82	20.83	23.74	23.67	**28.01**	23.74
Total	32.19	32.45	**39.17**	39.15	38.27	**39.17**

Tab. 4. Accuracy score for PoS tagging with 3-gram model (3-gram), enhanced with decapitalization (3 + D), decapitalization, and marks of token borders (3 + DB), the latter two and TF-IDF weighting (3 + DBT), or token length counter (3 + DBL), or scoring repeating symbols, and digraphs as one symbol (3 + DBR)

These results were still very low, so the next decision was to train a raw HMM model. The results are presented in table 5.

PoS	HMM
VERB	68.79
AUX	98.77
ADV	60.31
NOUN	99.04
PRON	86.51
CCONJ	99.65
ADP	98.7
ADJ	56.36
INTJ	91.53
SCONJ	70
DET	73.13
PROPN	87.33
NUM	92.56
X	20
Total	81.04

Tab. 5. Accuracy score for PoS tagging with raw HMM with Viterbi algorithm (HMM) model on the UD OCS test dataset module

As can be seen, both HMM and n-gram models tend to be biased towards a particular PoS. HMM is better at detecting nouns, though sometimes it fails in distinguishing them from other PoS. The same may be said of the n-gram model and verb. N-gram may be of better utility, while searching for particular PoS, like X, verb, and adjective. And, simultaneously, it may have a negative effect on the overall quality of tagging. So we built a set of hybrid HMM and n-gram models following the path paved by TreeTagger [35], but making them more adapted to the structure of the OCS data. The hybrids employed the following scheme. After the HMM model made preliminary tagging, the 3-gram model, with or without enhancements, checked tokens once again, assigning them to a preliminary defined PoS. These were adjective, adverb, verb, and X (non-word), for which the n-gram model chose the

correct tag with a higher probability than the HMM model. However, the first test defined that adverb tagging with the 3-gram model decreases overall efficiency, and the accuracy of the noun tag prediction decreases anyway. However, the 3-gram model additionally tags only adjective, verb, and X, total accuracy increases, at the cost of noun accuracy. A slight increase in efficiency, achieved by decapitalization, remained. In contrast, explicit representation of the token borders made a slight decrease in performance.

Almost each hybrid significantly increased accuracy of X detection. This is due to the fact that there is a very small number of X in the test dataset, and the difference is not more than 5 correctly assigned tags.

The final experiment included a regression model that learned to predict the correct tag on the basis of HMM and 3-gram prediction. Accuracy of this model is slightly unstable, due to the rounding implementation in Python. Having said that, it was still the baseline model which produced better results.

The results of experiments with hybrid models are given in table 6.

PoS	Baseline	HMM + 3	HMM + 3 – ADV	HMM + 3 + D – ADV	HMM + 3 + DB – ADV	HMM + 3 + DB – ADV + ETR
VERB	68.79	71.37	72.69	**72.82**	55.46	72.69
AUX	**98.77**	**98.77**	**98.77**	**98.77**	**98.77**	**98.77**
ADV	60.31	**80.06**	60.31	60.31	60.12	57.42
NOUN	**99.04**	94.66	98.08	98.08	96.16	98.2
PRON	**86.51**	72.79	**86.51**	**86.51**	86.13	73.32
CCONJ	**99.65**	**99.65**	**99.65**	**99.65**	**99.65**	6.83
ADP	**98.7**	98.18	98.18	98.18	82.47	**98.7**
ADJ	56.36	52.94	**56.9**	**56.9**	55.08	56.36
INTJ	**91.53**	**91.53**	**91.53**	**91.53**	**91.53**	**91.53**
SCONJ	**70**	63.04	**70**	**70**	**70**	19.57
DET	**73.13**	**73.13**	**73.13**	**73.13**	64.93	25.37
PROPN	**87.33**	86.3	**87.33**	**87.33**	83.22	87.33
NUM	**92.56**	91.74	91.74	91.74	90.08	63.64
X	20	**60**	**60**	**60**	**60**	20
Total	81.04	80.6	81.79	**81.82**	75.85	69.62

Tab. 6. Accuracy score for PoS tagging with HMM model (HMM), enhanced with 3-gram model (1), 3-gram model that does not make additional predictions for adverbs (2), the latter with decapitalization (3), and with explicit statement of token beginning and ending (4), the latter and the Extra Trees Regressor, picking the best possible option

For the following out-of-domain experiment run on the Kiev Folia dataset, architecture HMM + 3 + D – ADV was taken, because it demonstrated the best results in within-domain settings. Basic HMM model performed as a baseline in this case. The HMM + 3 + D – ADV model was additionally enhanced with rules that

help to define punctuation marks, digits, and fragmentary tokens. Enhanced HMM model performed better (50.93%) than the baseline one (32.64%) on Kiev Folia, as it had done previously on Codex Marianus. Quite expectedly, both models make significantly more mistakes than they did on the UD OCS dataset. Examples of Kiev Folia tokens tagging are provided in table 7.

Token	HMM tag	Enhanced model tag	Correct tag
Твоемоу 'yours-DAT/LOC'	NOUN	ADJ	ADJ
~*ИВ*~ '12'	NOUN	DIGIT	DIGIT
. '.'	NOUN	PUNCT	PUNCT
приведетъ 'lead-FUT(I)'	NOUN	VERB	VERB
присно 'always'	NOUN	VERB	ADV
приснодѣвѣ 'Mary, mother of Jesus-DAT/LOC'	NOUN	VERB	NOUN

Tab. 7. Examples of Kiev Folia dataset token tagging

5 ANALYSIS AND DISCUSSION

As one can see, the best models in our experiments achieve a rather non-impressive accuracy of 81% in in-domain settings and an accuracy of 51% in out-of-domain evaluation. In comparison, the UDpipe 2 neural tagger that employs the character level and multilingual BERT embeddings [29] achieves and accuracy of 97% and 50% respectively, being exposed to a larger amount of training data. The main explanation for such a dramatic drop is different letter distribution, different PoS tags distribution and even new letters and punctuation marks that have not been seen in training. Even the most frequent token, the coordinate conjunction 'and', is mostly presented as *u* in Codex Marianus and as *i* in Kiev Folia. We do not suggest the rule-based component of the tagger to be too specifically tuned to one particular text. Rather, the rules cover the PoS tags that are (almost) missing in the training set. The other part of work is done by the n-gram-based add-on that is based on assumption that there are morphemes, subtokens or other stable character combinations that can serve as a cue to the PoS identification. We believe that there is a space for future improvements of probabilistic models based on attention to most frequent character n-grams.

The analysis of the PoS confusion matrix on the in-domain and out-of-domain shows that verbs are frequently tagged as nouns and vice versa; adjectives with one-character endings are incorrectly labeled nouns (*петров-ъ* '*of Peter.POSS*', *мьног-ы* 'many'). At the same time, the closed-class PoS tags are identified mostly correctly in the in-domain test set, the only source of errors being the homonymy of prepositions and adverbs and conjunctions and adverbs or pronouns. In the out-of-domain test set, a lot of words from the closed-class PoS are erroneously tagged as nouns.

The disadvantage of the method is a partial loss of the context sensitivity. Thus, the noun *весь* 'village' is labeled determinant, as it has a homonymous and much more frequent reading 'all'. Analogically, единъ 'one' is labeled numeral when it refers to indefinite pronoun. Another known issue is nominalizations and other same-root and same-prefix words that get the tag of the most frequent word in their word formation family (usually, the basic word of the family). As a result, nouns that include n-gram *ѣл* in the root (*ѣлъ* 'verb', *ѣлсъ* 'voice') and nouns with the prefix *при* (*притъчи* 'parables', *пришелъца* 'newcomer') are incorrectly labeled verb. Apparently there should be found a sensible trade-off between the context sensitivity and subtoken recognition.

6 CONCLUSION

An HMM-based PoS tagger for OCS was developed, with some enhancements for both overall performance (n-gram models), and performance on specific PoS, such as digits and punctuation marks. The model achieves the accuracy score of 81% on the UD OCS dataset, and 51% on Kiev Folia dataset, which may satisfy the criterion of model scalability to out-of-domain data. HMM model, despite being in use since the early 1990s, was demonstrated to be still useful for a specific case of heterogeneous train and test data. What is more, the model seems to be operating better than pre-trained TreeTagger, that it was inspired by, and RFTagger [37] (14% overall accuracy on Kiev Folia).

The model is less accurate on UD OCS dataset than the UD multilingual models [29], however, it proves itself to be more useful for Kiev Folia dataset, due to the implementation of the rule-based systems. The results of out-of-domain evaluation yet again raise the question of how linguistically heterogeneous the OCS canon actually is. Apart from being practically useful for the OCS data annotation tasks, the cross-variant PoS tagging can provide actual insights into the scale of their difference.

The heterogeneity data case was a new challenge for OCS PoS tagging (comparing to [22] and [35]). And, with more texts being translated into a machine-readable form, and this model achieving 51% accuracy (which, as we hope, is not the best results that might be achieved), this challenge is to be faced in the future. Probably the following enhancements, like the ones that were conducted, are going to improve the overall results. The main aim here is to improve scalability of the model, not its efficiency for a single dataset.

References

[1] Behera, P. (2017). An Experiment with the CRF++ Parts of Speech (POS) Tagger for Odia Language in India, 17(1), pages 18–40.

[2] Loftsson, H. (2008). Tagging Icelandic text: A linguistic rule-based approach. Nordic Journal of Linguistics, 31(1), pages 47–72.

[3] Dandapat, S., Sarkar, S., and Basu, A. (2007). Automatic Part-of-Speech Tagging for Bengali: An Approach for Morphologically Rich Languages in a Poor Resource Scenario. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 221–224, Association for Computational Linguistics.
[4] Rajendran, S., and Krishnakumar. K. (2019). A Comprehensive Study of Shallow Parsing and Machine Translation in Malaylam. Coimbatore: Amrita Vishwa Vidyapeetham, 295 p.
[5] Uludoğan, G. (2018). HMM POS tagger. Accessible at: https://github.com/gokceuludogan/hmm-pos-tagger.
[6] Jurish, B. (2003). A Hybrid Approach to Part-of-Speech Tagging. Berlin: Berlin-Brandenburgishe Akademie der Wissenschaften, 2003, 27 p.
[7] (UD) UPOS tag set. Accessible at: https://universaldependencies.org/u/pos/.
[8] Mohamed Elhadj, Y. O. (2009). Statistical Part-of-Speech Tagger for Traditional Arabic Texts. Journal of Computer Science, 5(11), pages 794–800.
[9] Danso, S., and Lamb, W. (2014). Developing an Automatic Part-of-Speech Tagger for Scottish Gaelic. In Proceedings of the First Celtic Language Technology Workshop, pages 1–5, ACL.
[10] Mirzanezhad, Z., and Feizi-Derakhshi, M.-R. (2016). Using morphological analyzer to statistical POS Tagging on Persian Text. IJCSIS, 14(8), pages 1093–1103.
[11] Abumalloh, R. A., Al-Sarhan, H. M., Ibrahim, O. B., and Abu-Ulbeh, W. (2016). Arabic Part-of-Speech Tagging. Journal of Soft Computing and Decision Support Systems, 3(2), pages 45–52.
[12] Kumar, S. S., Kumar, M. A., and Soman, K. P. (2016). Experimental analysis of Malayalam PoS tagger using EPIC framework in Scala. ARPN Journal of Engineering and Applied Sciences, 11(13), pages 8017–8023.
[13] Gambäck, B., Olsson, F., Argaw, A. A., and Asker, L. (2009). Methods for Amharic Part-of-Speech Tagging. In Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages, pages 104–111, ACL.
[14] Saharia, N., Das, D., Sharma, S., and Kalita, J. (2009). Part of Speech Tagger for Assamese Text. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 33–36, World Scientific Publishing Co Pte Ltd.
[15] Reddy, S., and Sharoff, S. (2011). Cross Language POS Taggers (and other Tools) for Indian Languages: An Experiment with Kannada using Telugu Resources. In Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 11–19, Asian Federation of Natural Language Processing.
[16] Kann, K. et al. (2018). Character-level supervision for low-resource POS tagging. In Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP, pages 1–11, Association for Computational Linguistics.
[17] Straka, M., Strakova, J., and Hajič, J. (2019). Czech Text Processing with Contextual Embeddings: POS Tagging, Lemmatization, Parsing and NER. In Proceedings of 22nd International Conference "Text, Speech and Dialogue" 2019, pages 137–150, TSD.
[18] Hajič, J., and Hladká, B. (1998). Czech language processing, POS tagging. Accessible at: https://ufal.mff.cuni.cz/czech-tagging/HajicHladkaLREC1998.pdf.
[19] Korobov, M. (2015). Morphological Analyzer and Generator for Russian and Ukrainian Languages. Analysis of Images, Social Networks and Texts, pages 320–332.
[20] Bird, S., Loper, E., and Klein, E. (2009). Natural Language Processing with Python. O'Reilly Media Inc, 479 p.

[21] Segalovich, I. (2003). A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. In Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications, pages 273–280, MLMTA'03, June 23–26, 2003, Las Vegas, Nevada, USA.
[22] Eckhoff, H. M., and Berdicevskis, A. (2015). Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. Scripta & e-Scripta 2015, 14(15), pages 9–25.
[23] Pedrazzini, N., and Eckhoff, H. M. (2021). OldSlavNet: A scalable Early Slavic dependency parser trained on modern language data. Software Impacts, 8. Accessible at: https://www.sciencedirect.com/science/article/pii/S2665963821000117.
[24] TITUS. Accessible at: http://titus.uni-frankfurt.de/indexe.htm.
[25] Manuscript. Accessible at: http://manuscripts.ru/.
[26] Zeman, D., Nivre, J., Abrams, M. et al. (2020). Universal Dependencies 2.7. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Accessible at: http://hdl.handle.net/11234/1-3424.
[27] Haug, D. T. T., and Jøhndal, M. L. (2008). Creating a Parallel Treebank of the Old Indo-European Bible Translations. In Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008), pages 27–34, ACM, New York, NY.
[28] Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia. Accessible at: https://ufal.mff.cuni.cz/~straka/papers/2016-lrec_udpipe.pdf.
[29] Straka, M., Straková, J., and Hajič, J. (2019). Evaluating Contextualized Embeddings on 54 Languages in POS Tagging, Lemmatization and Dependency Parsing. In ArXiv.org Computing Research Repository, ISSN 2331-8422, 1904.02099.
[30] Kamphuis, J. (2020). Verbal Aspect in Old Church Slavonic: A Corpus-based Approach. Leiden: Brill, 329 p.
[31] Strobl, C., Malley, J., and Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. Psychological Methods, 14(4), pages 323–348.
[32] Kiev Folia. Accessible at: http://www.schaeken.nl/lu/research/online/editions/kievfol.html.
[33] Afanasev, I. (2020). Korpus staroslavianskogo iazyka: nedostaiushchee zveno v diakhronicheskoi slavistike. In Slavica iuvenum XXI: sbornik trudov mezhdunarodnoi nauchnoi konferentsii Slavica iuvenum 2020, March 31–April 1, 2020, pages 13–21, Ostravskii universitet, Ostrava.
[34] Project GitHub Repository. Accessible at: https://github.com/The-One-Who-Speaks-and-Depicts/hmm-pos-tagger.
[35] Helmut, S. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of International Conference on New Methods in Language Processing, pages 1–9, Manchester, UK. Accessible at: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf.
[36] Simov, K., Osenova, P., and Slavcheva, M. (2004). BTB-TR03: BulTreeBank Morphosyntactic Tagset. Accessible at: http://bultreebank.org/wp-content/uploads/2017/06/BTB-TR03.pdf.
[37] Schmid, H., and Laws, F. (2008). Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging. Accessible at: https://cis.lmu.de/~schmid/papers/Schmid-Laws.pdf.

sciendo
DOI 10.2478/jazcas-2021-0052

BUILDING AN EDUCATIONAL LANGUAGE PORTAL USING EXISTING DICTIONARY DATA

ANDREJ PERDIH – KOZMA AHAČIČ – JANOŠ JEŽOVNIK – DUŠA RACE
Fran Ramovš Institute of the Slovenian Language, Research Centre of the Slovenian Academy of Sciences and Arts, Ljubljana, Slovenia

PERDIH, Andrej – AHAČIČ, Kozma – JEŽOVNIK, Janoš – RACE, Duša: Building an educational language portal using existing dictionary data. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 568 – 578.

Abstract: The article presents the process of building the *Franček* Slovenian language portal aimed at primary- and secondary-school students. We discuss problems and solutions of linking and adapting existing non-pedagogical dictionaries for school use, while overcoming content and structural differences among the dictionaries. We also present some solutions within the process of adaptation to the online medium and visualisation adjustments for three age groups of school users with different content needs and levels of (meta)linguistic knowledge.

Keywords: pedagogical lexicography, language portal, Slovenian language, dictionary linking, children's dictionary

1 INTRODUCTION

Franček is an educational language portal for Slovenian aimed at primary- and secondary-school students. By building the portal we seek to provide a solution to a fundamental obstacle in the early use of dictionaries revealed by studies on the use of electronic resources in the Slovenian educational system ([1], [2]): their lack of adaptation to the users' age.

Since 2014, Slovenian primary- and secondary-school students have used online dictionaries of the Slovenian language only through the *Fran* portal. The *Fran* web portal combines thirty-eight dictionaries (with a total of 689,941 dictionary entries), a dialect atlas, and online language counselling and terminological counselling services, all searchable through a single search engine, displaying results from all the different sources at once ([3], [4]). It was set up in 2014 and quickly became popular in the Slovenian educational system: it is referred to in all recent Slovenian language textbooks, and its use is also promoted by the Slovenian National Education Institute.[1]

[1] E.g., https://www.zrss.si/objava/portal-fran-in-portal-francek-za-solsko-rabo. The *Fran* portal is exceptionally popular, with approximately 200,000 searches recorded daily at the time of writing.

Adaptation of dictionaries to better suit students was the main source of motivation behind designing the new *Franček* portal (https://www.francek.si), where dictionary material is displayed not by individual dictionary, but aggregated to provide wholesome information on individual words with regard to their meaning, synonymy, morphology, pronunciation, phraseology, dialect variation, history, and etymology.

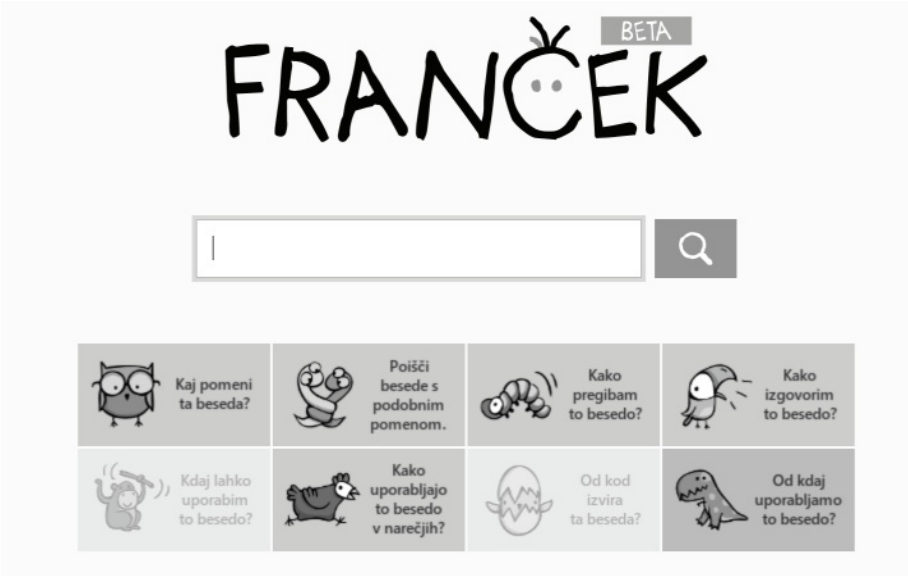


Fig. 1. The main site of the *Franček* portal

Franček combines materials from various lexicographic resources and presents the data in a simplified manner by providing answers to specific questions a student might ask.[2] Information on the data's primary source is clearly marked. This way, via the natural situation of learning about language (i.e., using questions and answers), students are gradually taught how to use and, more so, appropriately understand more complex dictionary content.[3]

Combining different dictionary databases in a single portal is an extremely demanding lexicographic challenge. For example, the label *pogovorno* 'colloquial' alone is used very differently in various Slovenian dictionaries; approaches to labelling

[2] Translations of questions presented in icons in Fig. 1 are as follows: "What does this word mean?"; "Find words with similar meaning."; "How is this word inflected?"; "How do I pronounce this word?"; "Which idioms does this word appear in?"; "How is this word used in dialects?"; "What is the origin of this word?" and "Since when has this word been used?".

[3] Every answer provided on the simplified *Franček* portal is linked to a dictionary entry on the *Fran* portal intended for experienced users.

parts of speech can differ significantly; dictionaries even differ in how they arrange specific headwords in different periods. How should a portal, then, be set up to suitably take account of the numerous differences between dictionaries and combine them reliably? How can conceptually diverse dictionaries be combined into a single format, while at the same time raising young users' awareness of the difference between them?

While designing *Franček*, these questions were addressed at the following three levels:

1) by linking databases to an initially constructed headword list,
2) by displaying dictionary data differently for each age group (i.e., grades 1 to 5, grades 6 to 9, and secondary school), and
3) by making content-related changes to the databases and preparing a suitable supplementary apparatus in the form of tooltips, altered metatext of individual dictionaries, by omitting certain less important information from dictionaries, and by providing links between dictionaries and a pedagogical grammatical description[4] (https://www.francek.si/kje-je-kaj-v-slovnici).

2 SEMI-AUTOMATED LINKING PROCESS

The portal is built around a central headword list (cf. 2.1), with eight modules linked thereto. These provide various linguistic information, namely on the words' meanings, synonyms, morphological paradigms, pronunciation, phraseology, dialect variation, etymology, and information on historical usage. The modules' contents are visualized from underlying databases based on the age group preselected by the user (cf 3.2).

The underlying databases were linked to the headword list using semi-automated linking processes (cf. 2.2). All automated processes were performed on dictionary databases in XML format using XSLT transformations. Manual linking was performed using the *iLex* dictionary writing system (DWS) [6]. Several parts of the data were exported to plain text or Excel files to manually select specific headword IDs to be explicitly included in the XSLT transformation processes.[5] New data, such as the new dictionary for school use *Šolski slovar slovenskega jezika* (ŠSSJ), was manually entered using the *iLex* DWS. Additionally, select dictionary data was modified or enriched to better the end-user experience (cf. 2.3).

2.1 Headword list

The headword list was established on the basis of two general monolingual dictionaries: *eSSKJ – Dictionary of the Slovenian Standard Language, 3rd Edition*

[4] The process of matching lexicographical data to appropriate descriptions within school grammars is presented in [5] in this publication.

[5] Original XML files did not follow a common standard schema, which had to be taken into account and made the preparation more time-consuming.

[7], and *Dictionary of the Slovenian Standard Language, 2nd Edition* (SSKJ2) [8]. While eSSKJ, the newer of the two dictionaries, was prioritized over SSKJ2, it currently contains a lot less than SSKJ2, which represents the vast majority of the headword list. Not all entries from SSKJ2 were accepted into the *Franček* database, as lexemes labelled as *zastarelo* 'obsolete' and *vulgarno* 'vulgar' were excluded. Certain types of SSKJ2 sublemmas were also included in the headword list, such as (non-)reflexive verb pairs and adverbs [9] (5481 out of 12,549 sublemmas, 43.68%). The vast majority of the inclusion/exclusion rules were used during the automatic headword-list building process.

2.2 Linking dictionary data to the headword list

Linking other resources to the headword list was first undertaken as an automated rule-based process, followed by a manual rechecking and linking process to address ambiguities and special cases. Entries were matched according to headwords and, where applicable, part-of-speech data and stress placement. Some caution had to be exercised with regard to POS labels due to different underlying grammatical theories used in different dictionaries; the differences had to be reconciled and the data normalized. While such differences come as no surprise in the case of historical dictionaries, the same issue arose also in the process of linking the *Synonym Dictionary of Slovenian Language* (SSSJ; [10], [11]), even though it is based on SSKJ (1st edition) [12].[6] Stress placement was sometimes also used to automatically differentiate between homographs. Perfect homonyms had to be disambiguated and linked manually.

Links were established at headword level only. We did not attempt to systematically link information at sense level, nor were sense-level gaps filled if the data was available. Consequently, it is possible that a sense not covered in the semantic module may appear in other modules.

This is most evident in the case of the dialect module, since the main source of dialect lexical data used, the *Slovenian Linguistic Atlas* (SLA; [13], [14]), is primarily onomasiological in nature (as opposed to all the other dictionaries, which are classic semasiological dictionaries). The dialect module is divided into two sections: the onomasiological section lists dialect lexemes denoting the meaning of Standard Slovenian lexical forms (i.e. it seeks to provide answers to the question "which words are used to describe this concept and in which dialects?"), and the semasiological section provides alternative dialect senses of Standard Slovenian lexical forms (i.e. "which concepts does this word (also) denote and in which dialects?") [15].

[6] The main reasons for discrepancies are the treatment of the predicative (*povedkovnik*) as a standalone POS (i.e. nouns, adjectives, and adverbs can be interpreted as predicatives depending on their syntactic function and treated as separate lexemes), and the treatment of qualitative and classifying adjectives as separate lexical units (e.g. *zelen* 'green – qualitative adjective': *zeleni* 'green – classifying adjective'); SSKJ does not distinguish between these categories.

Semantic disambiguation proved to be most problematic in the case of the historical module, as the underlying dictionaries describe different lexical systems, dating from mid-16th to late 19th century. Lexemes that semantically greatly differed from their modern Slovenian counterparts, or those whose senses are no longer attested, were manually excluded; e.g. in the case of homonyms *moka* 'flour' and *moka* 'anguish, torment', the latter was excluded, as it had already fallen out of use by the end of the 16th century to eventually be replaced by its Slavic cognate *muka*. Furthermore, differences in orthographic principles among dictionaries had to be taken into account. In cases where the orthographic forms differed from the modern Slovenian ones, the closest form was chosen; e.g. deverbal nomina agentis such as *bravec* 'reader' or 'gatherer', *igravec* 'player' or 'actor', *plezavec* 'climber' etc. were linked to their modern Slovenian counterparts *bralec*, *igralec*, *plezalec* etc. This principle was also adhered to in a limited scope in the case of non-systemic orthographic forms, e.g. *ambašador* was linked to the headword *ambasador* 'ambassador, envoy' (while *bašador*, also attested in the same resource, was excluded).

2.3 Modifications and enhancement of dictionary data

Some data was significantly altered prior to inclusion in the *Franček* database. This was mostly due to the fact that the original resources were less suitable for educational use and thus required simplification. Such was the case with the dialect module, where the main source was the index of the SLA atlas, and a subsection of the historical module, where the main source was a register of all lexemes attested in 16th century Slovenian texts [16]. The latter was used to create a database of the earliest attestations of lexemes in written form.[7]

In SSKJ2, labels pertaining to all or to the majority of senses are presented in the head of the dictionary entry, i.e. at the entry level. As presentation of data from the head is limited, and to give more understandable information to the end-user in the semantic module itself, entry-level labels were transferred to sense level. While the process was automated, complex rule refinement was needed to account for cases where entry-level labels needed to be omitted, usually when the entry-level and sense-level labels belonged to the same type (e.g. stylistic labels or register labels). In some cases, the entry-level labels needed to be placed after the sense-level label to meet the sorting rules in the dictionary (e.g. *ekspresivno + pogovorno* 'expressive, colloquial' was changed to *pogovorno + ekspresivno*). Some other exceptions also had to be taken into account, as it was not possible to combine the *starinsko* 'archaic'[8] label with the majority of other label types. Special treatment was necessary also

[7] The database excludes a relatively small number of lexemes attested in earlier Slovenian manuscripts due to the unavailability of data in digital form.

[8] The dictionary differentiates between *obsolete* and *archaic* lexis.

with regard to grammatical labels as some combinations are possible, while some labels are mutually exclusive.

Cross-referencing (excluding referential definitions in the semantic module) was reduced as much as possible by inserting the target content in place of the reference's origin, most evidently so in the case of the etymological module.

In the cases of homographs and homonyms, indicators were created to help young users disambiguate among them. While the basic distinction can be done by indicating POS information (1242 entries, automated process), 3213 semantic indicators were also added (out of 4384 total manually added to the underlying database).[9] If neither POS nor semantic indicators could be used for disambiguation, morphological or stress-placement indicators were provided (35 entries).

3 CUSTOMISED DATA VISUALISATION

The *Franček* portal is aimed at students of three age-groups:

1) 1st to 5th grade of primary school,
2) 6th to 9th grade of primary school,
3) secondary school.

ŠSSJ ([17], [18]) seeks to fulfill the needs of the first age group; its extent and concept are adapted to the children's abilities and needs based on the curriculum. This dictionary contains 2000 entries pertaining to basic vocabulary and is displayed in the semantic module.[10] Even though their use in the educational process is expected and planned, other resources used during the creation of *Franček* are not primarily intended for school use.[11] Their excessive complexity, especially in the case of the SSKJ, is well documented ([22], [23]) and strengthens the assumption among students, teachers, and other dictionary users that successful use of dictionaries is something that has to be learned, and that one should practice using dictionaries. As already noted by Tarp [24], a dictionary is not merely a list or a language database; rather, it is primarily a practical tool for language use, which retrieves information from a database as required by the user. The aim of *Franček* is, therefore, to provide language data in a way that will be useful (selection of relevant content) and understandable (adaptive visualization) to students. Since we used existing language resources that had not been created with students in mind, and some of which had not been primarily made for the web, we

[9] The difference stems from the fact that obsolete homonyms were omitted from the *Franček* database.

[10] In other entries, simplified content of SSKJ2 and eSSKJ is displayed.

[11] The 2018 curriculum for Slovene lessons in primary school envisages the use of dictionaries mainly from the 5th grade onwards (in teaching materials, students are most often directed to SSKJ and the *Slovenian Normative Guide* [19]); even before that age, children are expected to be able to at least identify the meanings of words ([20], [21]).

had to find solutions for visualisation of language material that would meet the users' requirements.

Franček features two types of customised visualisation of language data:

1) adjustments due to transfer to the online medium and to the portal design,
2) adjustments due to changed target users.

3.1 Adjustments to the online medium

While eSSKJ and ŠSSJ are primarily online dictionaries, other resources were made as print dictionaries; content is thus structured in a condensed manner due to limited space, which is reflected in the implicit presentation of information with different types of font, abbreviations, symbols, etc. All abbreviations, labels, and symbols were made explicit on *Franček*, e.g. instead of introductory symbols (such as ♦ for the terminological section in SSKJ2) the content is clearly explained (e.g. "*This word is a professional term*"). Labels have not only been fully spelled out (e.g. *pog.* as *pogovorno* 'colloquial') but also explained in tooltips ("A word, multi-word unit, or sense used especially in everyday and less formal communication") and linked to appropriate chapters in school grammars [5]. Visual and audio materials were added. For structuring the content and navigating the portal, standard icons (e.g. the microphone icon indicates the possibility of recording; the map pointer icon prompts users to view a map, etc.) and established web conventions are adhered to (e.g. use of tooltips, links to detailed information, etc.). Styles and colours are consistent throughout the portal (e.g. illustrative material is always green, clickable content is blue, sense numbering is highlighted in blue etc.).

3.2 Adjustments for new target users

Specific age-group requirements and levels of linguistic and metalinguistic knowledge were considered when adapting the display of dictionary data.

A user in the **lowest age group** is, therefore, not overburdened with the dictionary metalanguage and microstructure. Illustrated icons with simple explanations in tooltips are used to provide information on certain stylistic characteristics and grammatical categories (e.g. countable nouns are represented by icons of dice with one (singular), two (dual), and three (plural) dots; outdated synonyms are introduced with an icon of an old man, etc.). In other cases, helping hints were added, e.g. appropriate question words were added next to names of grammatical cases in the morphological module. In this age group, content is limited to semantic, synonymic, morphological, and pronunciation modules.

Visualisation for students of **6th to 9th grade** takes into account that some users in this age group are already familiar with basic metalanguage and use dictionaries for writing. Parts of the dictionary microstructure are explicitly marked (definition, examples, typical constructions, variants). The content is more extensive, as the phraseological, dialect, etymological, and historical modules are included. However,

the number of listed terms, multi-word units, and their variants is limited. Illustrated icons are still used at this stage to symbolize grammatical categories and stylistic characteristics for ease of memorisation, while additional aids (e.g. question words, short explanations of more complex grammatical categories, etc.) have been moved to tooltips.

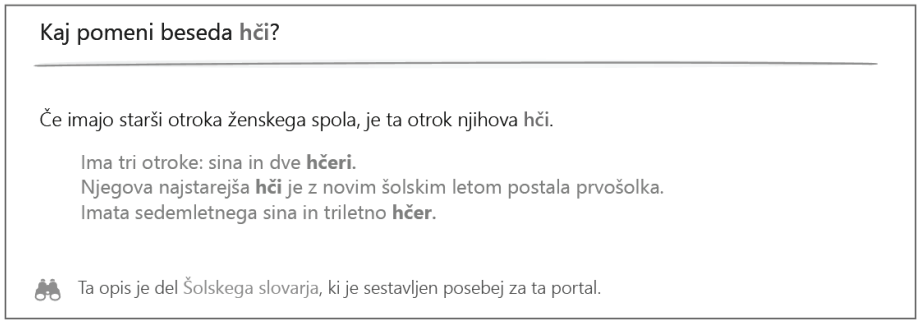


Fig. 2. Visualisation of the semantic module for students of 1st to 5th grade

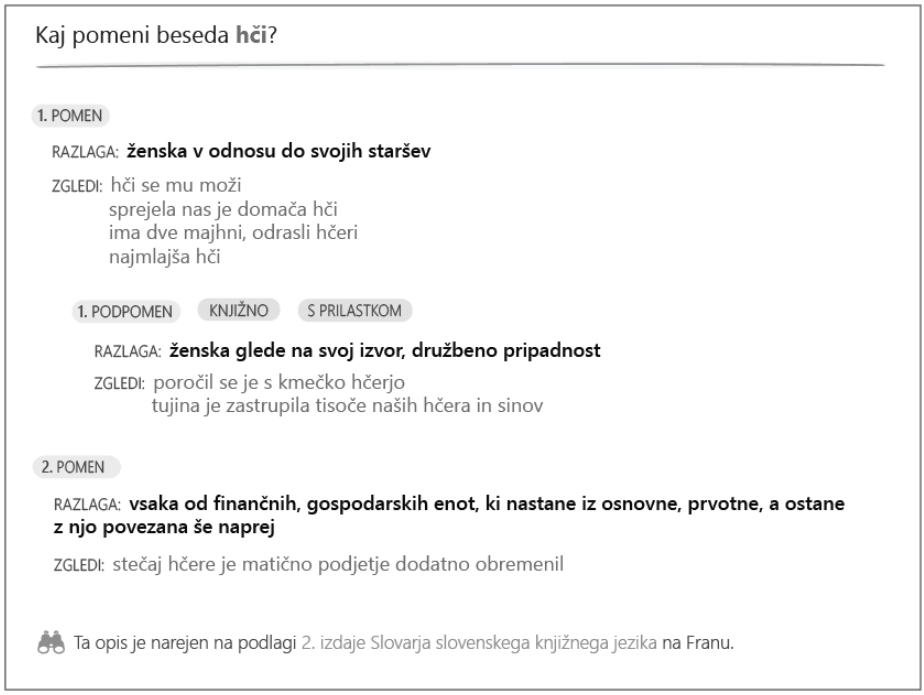


Fig. 3. Visualisation of the semantic module for students of 6th to 9th grade

Visualisation of dictionary content for **secondary-school students** relies on the fact that the users are already familiar with the microstructures of various dictionaries; therefore, dictionary metalanguage is not explicitly presented or graphically illustrated.

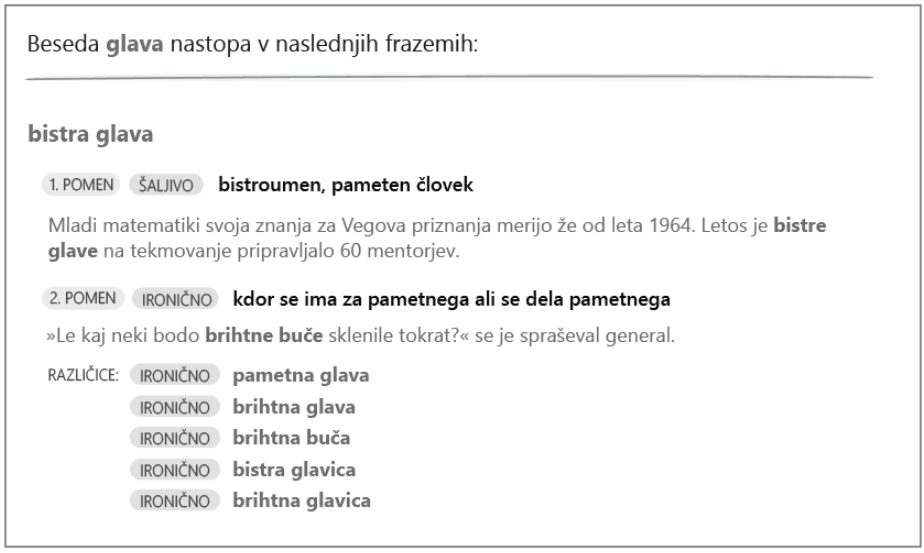


Fig. 4. Visualisation of a phraseological module for secondary-school students

Although visualisation (and content complexity) is only a small step away from that on the *Fran* portal (especially in eSSKJ), explanations in tooltips are still a notable advantage in comparison (e.g. explanation of animacy in the morphological section, links to descriptions within the school grammar, explanations of labels, etc.) and enable proper interpretation of data without detailed knowledge of the concept of the source dictionary.

4 CONCLUSION

Franček, the new educational Slovenian language portal, was built to fill the gap in Slovenian linguistic resources for educational purposes. Lexicographic data on the portal was adapted and linked from existing non-pedagogical dictionaries, while new data was also prepared specifically for this purpose. The lexicographic content is presented from the point of view of individual words, creating a single lexicographic resource. The data is organized in eight modules: semantic, synonymic, morphological, pronunciation, phraseological, dialect, historical, and etymological. Content and visualisation are adjusted to the online medium and adapted to three age groups of

users: primary-school students from 1st to 5th grade and 6th to 9th grade students, and secondary-school students. Additionally, lexicographic data presents a part of a wider ecosystem of linked lexicographic, grammar, and language counselling data.

ACKNOWLEDGEMENTS

This article was produced as part of the project *Portal Franček, Jezikovna svetovalnica za učitelje slovenščine in Šolski slovar slovenskega jezika* (The Franček Portal, the Language Counselling Service for Teachers of Slovenian, and the School Dictionary of Slovenian) co-funded by the Republic of Slovenia and the European Social Fund; part of the research for the project was conducted within the Slovenian Research Agency's P6-0038 program group The Slovenian Language in Synchronic and Diachronic Development.

References

[1] Kosem, I., Stritar, M., Može, S., Zwitter Vitez, A., Arhar Holdt, Š., and Rozman, T. (2012). Analiza jezikovnih težav učencev: korpusni pristop. Ljubljana: Trojina, zavod za uporabno humanistiko, 132 p.

[2] Rozman, T., Krapš Vodopivec, I., Stritar, M., and Kosem, I. (2020). Empirični pogled na pouk slovenskega jezika. Ljubljana: Znanstvena založba Filozofske fakultete, 183 p. Accessible at: https://e-knjige.ff.uni-lj.si.

[3] Ahačič, K., Ledinek, N., and Perdih, A. (2015). Fran: the next generation Slovenian dictionary portal. In Natural language processing, corpus linguistics, lexicography: Proceedings, Eighth International Conference, Bratislava, Slovakia, 21–22 October 2015, pages 9–16, RAM-Verlag. Accessible at: https://korpus.sk.

[4] Perdih, A. (2020). Portal Fran: od začetkov do danes. Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje, 46(2), pages 997–1018.

[5] Ahačič, K., Ledinek, N., and Petric Žižić, Š. (manuscript submitted for publication). Presenting and linking grammatical data on the Franček educational language portal.

[6] Erlandsen, J. (2010). iLEX, a general system for traditional dictionaries on paper and adaptive electronic lexical resources. In Proceedings of the XIV EURALEX International Congress. 6–10 July 2010, page 306, Leeuwarden/Ljouwert: Fryske Akademy – Afûk. Accessible at: https://euralex.org.

[7] eSSKJ: Slovar slovenskega knjižnega jezika (2016–). Accessible at: https://www.fran.si.

[8] Slovar slovenskega knjižnega jezika, druga, dopolnjena in deloma prenovljena izdaja (2014). Accessible at: https://www.fran.si.

[9] Perdih, A. (manuscript submitted for publication). Učenje o slovarjih v šoli: portal Franček kot most med splošno in pedagoško leksikografijo.

[10] Snoj, J., Ahlin, M., Lazar, B., and Praznik, Z. Sinonimni slovar slovenskega jezika (2018 [2016]). Accessible at: https://www.fran.si.

[11] Snoj, J. (2019). Leksikalna sinonimija v Sinonimnem slovarju slovenskega jezika. Lingua Slovenica 14. Ljubljana: Založba ZRC, 316 p.

[12] Slovar slovenskega knjižnega jezika (2014 [1970–1991]). Accessible at: https://www.fran.si.
[13] Slovenski lingvistični atlas 1 (2014 [2011]). Accessible at: https://www.fran.si.
[14] Slovenski lingvistični atlas 2 (2016). Accessible at: https://www.fran.si.
[15] Ježovnik, J., Kenda-Jež, K., and Škofic, J. (2020). Reduce, Reuse, Recycle: Adaptation of Scientific Dialect Data for Use in a Language Portal for School children. In Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. I., pages 31–37, Democritus University of Thrace. Accessible at: https://euralex2020.gr.
[16] Besedje slovenskega knjižnega jezika 16. stoletja (2014 [2011]). Accessible at: www.fran.si.
[17] Godec Soršak, L. (2015). Slovenski otroški šolski slovar. In Slovnica in slovar – aktualni jezikovni opis (1. del), Obdobja 34, pages 243–250. Ljubljana: Znanstvena založba Filozofske fakultete. Accessible at: https://centerslo.si/simpozij-obdobja/zborniki/.
[18] Petric Žižić, Š. (2020). Tipologija razlag v Šolskem slovarju slovenskega jezika. Slavistična revija, 68(3), pages 391–409.
[19] Slovenski pravopis (2014 [2001]). Accessible at: https://www.fran.si.
[20] Godec Soršak, L. (2020). Spodbujanje rabe slovarja v učnem gradivu za slovenski jezik v 1. in 2. vzgojno-izobraževalnem obdobju. In Slovenski jezik in književnost v srednjeevropskem prostoru: Zbornik SDS 30, pages 235–244. Ljubljana: Zveza društev SDS.
[21] Petric Žižić, Š. (2020). Usvajanje besedoslovne jezikovne ravnine in raba slovarjev pri pouku slovenščine v osnovni šoli (pregled učnega gradiva za tretje vzgojno-izobraževalno obdobje). In Slovenski jezik in književnost v srednjeevropskem prostoru: Zbornik SDS 30, pages 245–253. Ljubljana: Zveza društev SDS.
[22] Vrbinc, M. (2004). An empirical study of dictionary use: the case of Slovenia. In ELOPE: English Language Overseas Perspectives and Enquiries, 2(1–2), pages 97–106. Ljubljana: Ljubljana University Press, Faculty of Arts.
[23] Rozman, T. (2010). Vloga enojezičnega slovarja slovenščine pri razvoju jezikovne zmožnosti (PhD thesis). Ljubljana: Filozofska fakulteta, 359 p.
[24] Tarp, S. (2014). Detecting user needs for new online dictionary projects: Business as usual, user research or …? In Research into dictionary use: Wörterbuchbenutzungsforschung. 5. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie“, pages 16–26. Mannheim: Institut für Deutsche Sprache.

sciendo
DOI 10.2478/jazcas-2021-0053

StressDat – DATABASE OF SPEECH UNDER STRESS IN SLOVAK

RÓBERT SABO[1] – ŠTEFAN BEŇUŠ[1,2] – MARIAN TRNKA[1] – MARIAN RITOMSKÝ[1] – MILAN RUSKO[1] – MEILIN SCHAPER[3] – JAKUB SZABO[4]
[1] Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia
[2] Constantine the Philosopher University, Nitra, Slovakia
[3] Institute of Flight Guidance, German Aerospace Center, Braunschweig, Germany
[4] Institute of Molecular Biomedicine, Faculty of Medicine, Comenius University, Bratislava, Slovakia

SABO, Róbert – BEŇUŠ, Štefan – TRNKA, Marian – RITOMSKÝ, Marian – RUSKO, Milan – SCHAPER, Meilin – SZABO, Jakub: StressDat – Database of speech under stress in Slovak. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 579 – 589.

Abstract: The paper describes methodology for creating a Slovak database of speech under stress and pilot observations. While the relationship between stress and speech characteristics can be utilized in a wide domain of speech technology applications, its research suffers from the lack of suitable databases, particularly in conversational speech. We propose a novel procedure to record acted speech in the home of actors and using their own smartphones. We describe both the collection of speech material under three levels of stress and the subsequent annotation of stress levels in this material. First observations suggest a reasonable inter-annotator agreement, as well as interesting avenues for the relationship between the intended stress levels and those perceived in speech.

Keywords: speech database, speech under stress, stress annotation, inter-annotator agreement

1 INTRODUCTION

Areas of potential application for automatic speech technologies have been rapidly growing in the last decades. Despite great advances in statistical modelling and processing of speech, current state-of-the-art solutions still commonly use dedicated speech databases for specific domains of application. For example, in order to detect alcohol intoxication from speech, a system requires a specific database with speech under alcohol intoxication. In short, in order to be able to advance progress in the field of research and training of tools for automatic identification of speech expressiveness, it is necessary to build speech databases that will contain such speech expressions.

One such specific domain with great potential for making real world applications more effective and reliable is modelling stress based on speech characteristics. Understanding speech characteristics when people are under stress might help in mitigating stress-related problems and dangers in various situations,

such as air traffic control or crisis situations. This domain falls within a larger research area of modelling speech emotion and expressiveness ([1], [2], 3]). Recent years have witnessed immense progress in this modelling. This partly also results in the creation of and public access to several speech databases, which are specifically dedicated to research on emotions and expressiveness in speech ([4], [5]).

However, there is still a lack of suitable databases in speech under stress. The best-known database is SUSAS (Speech Under Simulated and Actual Stress [6]) that consists of four domains, encompassing a wide variety of stresses and emotions. It contains 32 speakers who have uttered more than 16,000 utterances. However, there are also limitations using SUSAS for training of tools for automatic identification of speech under stress and in performing acoustic analyses. Most sentences are one-word or two-word commands, there is often significant background noise, and the recordings have a low sampling frequency of 8 kHz.

Due to the lack of available corpora of naturalistic continuous speech containing stress level annotation, we decided to design and develop a dedicated Slovak database (StressDat) that would facilitate modelling speech under stress. The current paper describes our approach to the design, data collection, and processing. Specifically, section 2 describes the stimuli, section 3 speech recording, and section 4 the annotation of stress in the recordings. We also present first observations, particularly regarding the relationship between the intended and the perceived levels of stress in section 5. Section 6 presents further lines/directions of research and the conclusion.

2 StressDat STIMULI

2.1 Acted out vs. spontaneous speech

The first decision that had to be made was whether to use acted out or real speech. Naturally, spontaneous speech reflects the state of mind of the speaker in the best possible way. Additionally, some of the physical signs of stress are not possible to act out and, thus, their manifestation in speech is very difficult to imitate even for experienced actors.

On the other hand, a database of acted out speech in this domain has multiple advantages over spontaneous speech. First, using acted out speech has been a standard in the research of emotional and expressive speech for decades ([7], [8]). The main upshot is to have more control over the speech material and the level of stress. Spontaneous speech would have to be selected from various heterogeneous sources, situations, or recording conditions in a time-consuming effort. Acted out speech allows for recordings of more speakers in less time. Second, we can also control the situational context and limit the potential effects of a particular situation on the speech material in spontaneous speech. Rather, we can create specific stress situations that best reflect the intended use and coverage of the speech data. Third, the goal is to achieve balance in the amount of material for different levels of stress in speech, which would be immensely laborious to achieve with samples from speech under real stress. Fourth,

the quality of the acoustic signal is also important and it is essential to have comparable recordings regarding the same quality and the same amount of background noise. Finally, exposing participants to real stress might be ethically problematic while acting out a stressing situation does not pose these problems.

In addition to the above advantages of using acted out speech, we also tested the physiological indicators of stress when actors imitate a stressful situation. We asked an actor to read several sentences in neutral stress level and then imitate low, medium and high levels of stress. Our preliminary results suggest that even when an actor acts out a stressing scenario, physiological symptoms of real stress, such as increased skin conductivity and increased heart rate, can be observed. For each stress level, we measured average beats per minutes (BPM): neutral – 65.8, low – 65.4, medium – 68.05, high 72.7). These symptoms are, of course, weaker than when a person is actually exposed to a stressful situation. Figure 1 shows the temporal variability of skin conductivity G and heart rate HR during four acted out levels of stress (neutral, low, medium, high). For each of the three non-neutral stress levels, we can observe a typical step increase in both physiological indicators of stress, followed by their gradual decrease, indicating how the actor copes with this stress.

Hence, acted out speech is less natural than real speech, but a preferred option due to multiple practical considerations. Moreover, acted out speech is also linked to physiological indicators of stress in a similar way as real stress is expected to be linked to the same.

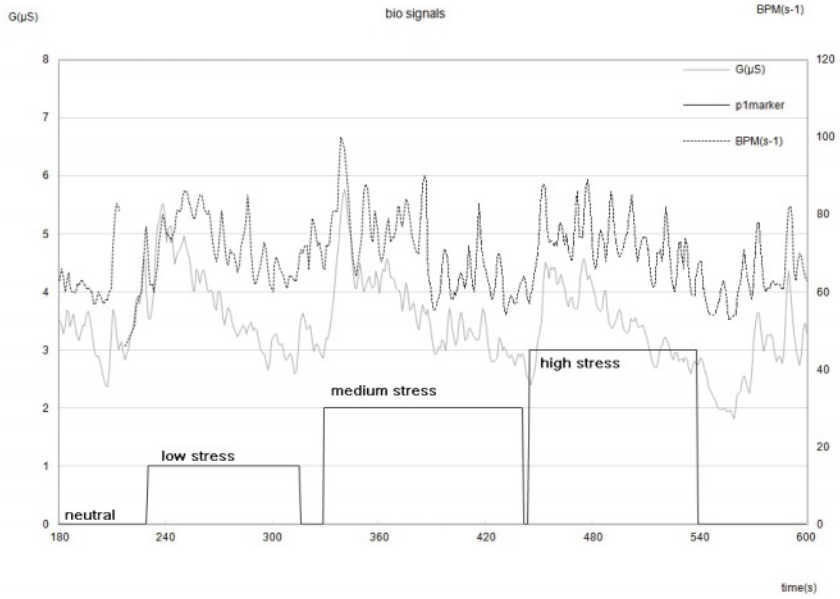


Fig. 1. Skin conductivity G (gray) and heart rate HR (dotted line) of the speaker acting out three levels of increased stress. Black solid line presents the stress level intended by the speaker

2.2 Stress levels

The decision to use acted out speech allows for some control over the levels of stress in the recording. Ideally, the database should cover as many levels of stress as people are able to produce systematically in a non-overlapping manner. We hypothesized that it would be two or three such levels.

This hypothesis is motivated in part by our prior research on developing a database of speech in crisis situations [9]. Naive subjects were able to produce three levels of activation corresponding to three levels of danger in a situation and the acoustic-prosodic characteristic in these three levels showed sufficient separation and reasonably acceptable overlap.

Additionally, in the pilot data shown in Figure 1, it can be observed that despite the typical step increase and gradual lowering of the two physiological indicators present for all three (low, medium, high) stress levels, the actual difference between the medium and high levels is non-existent in this particular speaker. Hence, while two levels of stress seem to be easily induced also in terms of physical indicators, three levels might be questionable and four rather unlikely.

Hence, given these considerations, and the fact that we planned to involve professional actors used to acting out various states of mind, we opted for three levels of stress (neutral, low, high).

2.3 Linguistic material

In order to obtain phonetically comparable recordings enabling measurements focused on acoustic and phonetic signs of stress level in speech, we created situations and associated textual material for acting out all three levels of stress (neutral, low stress and high stress). To capture the heterogeneity of stress and its manifestations in speech, we modeled 12 stressful situations. These could be grouped into three broad categories as follows: a) Threat of losing control over the situation, b) Psycho-social stress c) Threat to life/health or of an injury of self or the close ones.

Table 1 shows brief descriptions of these situations organized along three categories.

Category	Nr.	Stress level	Description
Threat of losing control over the situation	1	neutral, medium, high	As an airline pilot you need to make an emergency landing.
	2	neutral, medium, high	Navigating a plane at the airport during very bad weather.
	3	neutral, medium, high	As a pilot, you need an undisciplined passenger to comply with the ban on using laptops during takeoff/landing.
	4	neutral, medium, high	As a firefighter coordinator, you organize firefighting in a burning building.

Category	Nr.	Stress level	Description
Psycho-social stress	5	neutral, medium, high	As a parent, you have to organize the morning routine for your kids before leaving for school.
	6	neutral, medium, high	You and your colleague are making last-minute changes to an important presentation with a colleague.
	7	neutral, medium, high	As a passenger, you need information on train departures urgently.
Threat of to life/ health injury of self/close ones	8	neutral, medium, high	You are calling an ambulance for your father who has suffered a stroke.
	9	neutral, medium, high	You are trying to pacify your drunk brother who is trying to forcefully enter your flat.
	10	neutral, medium, high	You are calling the police to resolve the situation with your drunken brother above.
	11	neutral, medium, high	As a pilot, you organize evacuation from a burning aircraft.
	12	neutral, medium, high	You are reporting an insurance event after a car accident by phone.
Neutral	13	neutral	You are talking about school with your son.
	14	neutral	You are buying shoes.
	15	neutral	You are teaching students at school.
	16	neutral	You are reading a text to a colleague.

Tab. 1. Description of situations in the database

The goal was to achieve balance among different factors that might cause stress and different linguistic material for these factors. Each situation included between 10 to 13 sentences naturally expected in the given context and the sentences were created in a way that makes them appropriate for each stress level: neutral, under low stress, and under high stress. For example, a sample of sentences in situation #2 from Table 1 is shown in the following bulleted list.

- Gama 2305, it is important to quickly finish fueling the aircraft.
- Please speed up the loading of luggage, it is necessary to finish the loading of luggage as soon as possible.
- Runway number seven is not cleared from snow. Runway seven needs to be cleared. I repeat, runway seven needs to be cleared.
- The weather is getting worse, you need to take off immediately.
- ...

In addition to the 12 emotionally charged situations, we also included four 4 emotionally neutral situations with sentences corresponding to the neutral level of

stress only. These sentences were included since we cannot rule out the effect of text expressiveness on the neutral level of stress in the 12 expressive situations. These are shown at the bottom of Table 1. This inclusion will allow for testing the effect of linguistic material on the acoustic-prosodic rendering under the intended neutral level of stress (sentences corresponding to situations 1–12 vs. 13–16).

3 StressDat RECORDING

3.1 Speaker selection

To maintain the highest possible naturalness of elicited speech in the database, professional actors were recruited. The current pandemic situation facilitated recruiting of the actors since many of them experienced decreased demands on their time.

Currently the database includes 30 speakers (16 females, 14 males) who provided their recordings in exchange for payment. 20 speakers recorded the full battery of 16 situations in Table 1 and 10 speakers recorded 10 situations in 3 levels and 2 neutral situations.

3.2 Recording procedure

We needed to create a database of speech under stress at a time when people's face-to-face interactions were limited by the corona virus pandemic. For this reason, a novel procedure of database creation was developed. This allowed to not only achieve the required speech-under-stress recordings, but also to limit physical contact normally required in traditional speech elicitation protocols.

The goal was to utilize, and adjust if needed, the actors´ home environment and their own smartphones. We instructed the actors to select a room with the smallest possible reverberations, for example having as few bare walls and surfaces as possible, and make adjustments to further improve the acoustic environment, such as spreading the curtains, opening wardrobes, or covering sharp furniture edges. Additionally, instructions for positioning their smartphones during the recording were also given to ensure as comparable a recording environment across the speakers as possible.

The core of the instruction was to describe the three stress levels and facilitate actors' getting into the character. This was achieved in two ways. First, there were instructions regarding the three stress levels generally. For the neutral level, we asked them to imagine that they are completely calm, they navigate the situation with sufficient perspective, and that the situation does not affect their mental state in any adverse way. For the medium stress level, we asked them to imagine that they are under stress, that the situation is serious and its resolution should be done with care but assertively. For the high level of stress, we explained that they are under very high stress, that the situation is extremely critical and almost impossible to manage, and that they have to resolve it immediately.

Second, inspired by [10] for each situation, we specifically described a) the character (e.g., a parent of two schoolchildren that are difficult to manage), b) the situation (e.g., a Monday morning, an important meeting with grave consequences at work and parental duties involving the morning routine), c) the stressing factor (the family overslept and the kids are not cooperative), d) the goal (e.g., to manage to send the kids off to school and come to work on time), and e) the approach corresponding to three levels of stress (e.g., calmness usually works best with the kids (neutral), radio announces traffic jams and you need to be very efficient and effective with the kids (medium), you are very late, kid still doesn't behave, the situation is critical (high)).

At first, each actor recorded their first attempt at the two situations. We assessed both the acoustic quality of the recording and the differentiation of speech under the three stress levels. If adjustments were deemed necessary, they were communicated to the actor. The actors then proceeded with recording the full set of the situations.

4 StressDat ANNOTATION

After speech elicitation and processing, the annotation of the perceived level of stress in the recorded sentences was organized. A web-based speech stress assessment tool "Stress Thermometer" was designed based on the Subjective Units of Distress Scale [11], which allows the annotator to listen to the utterance and to assign a perceived stress level according to the instructions (see Figure 2). The visual representation of the thermometer was adopted from [12].

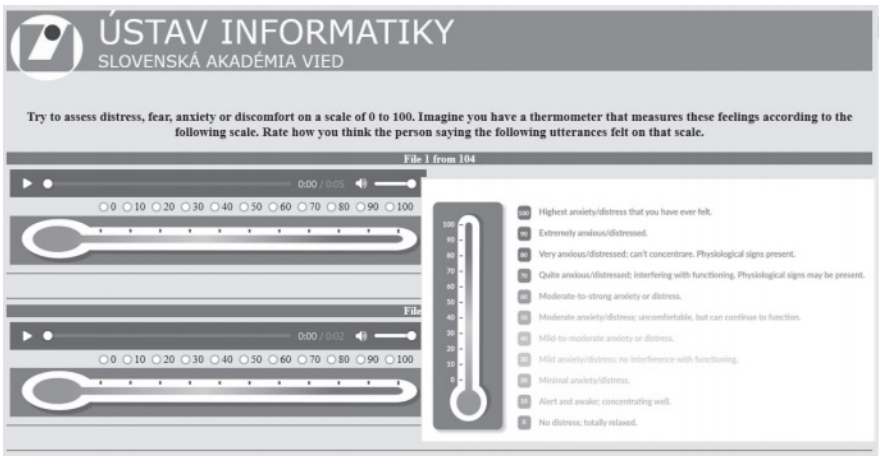


Fig. 2. The graphical user interface of the "Stress Thermometer" tool that allows the annotator to listen to the utterance and to assign a perceived stress level according to the verbal descriptions in the rightward panel [13]

Five annotators listened to each utterance and rated it on a discrete eleven-point scale according to the following instruction: “Try to assess distress, fear, anxiety, or discomfort on a scale from 0 to 100. Imagine you have a thermometer that measures these feelings on such scale. Rate how you think the person saying the following utterances felt on that scale.” Each utterance can thus be characterized with the mean of the values of the perceived stress level assigned by the annotators. The ratings of perceived stress can, therefore, reach real number values in the interval 0 to 100 in steps of 10. This allows regression to be used in stress assessment instead of classification. To limit the influence of the speaker, the annotators evaluated sentences from different speakers in random order. During the evaluation, each annotator had a different order of sentences in order to minimize the influence of the previously heard sentences on the evaluation.

Of the material material recorded by 30 actors, two thirds have been fully annotated and the rest is currently approaching completion.

5 PILOT OBSERVATIONS

5.1 Annotation normalization

It is common in annotating tasks using a scale that annotators use the scale in different ranges and variances. To normalize for this variability, we use z-score normalization [14] by annotator.

Figure 3 shows that normalizing annotations makes sense. Consider the neutral (Level 1) stress for raters a1–2 vs. a3–4. It is clear that the ratings of a1–2 are shifted lower compared to a3–4 in all three levels. Hence, the stress level was perceived similarly, only the first group used the lower range of the scale compared to the second group. This similarity among raters is reflected in the right panel after normalization. The figure also shows consistent and robust separation among the three stress levels in the annotations.

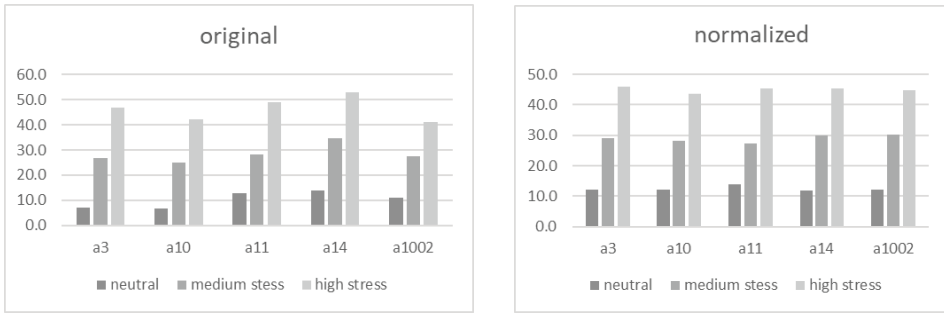


Fig. 3. Mean stress assessments for the three stress levels (neutral, medium, high) for five annotators (a1–a5) before (left) and after (right) normalization

5.2 Inter-annotator agreement

To find out the degree of agreement among all annotators using the 11-point scale, we calculated Fleiss' kappa [15] for the original and the normalized assessments, see the mean values for all utterances in Table 2. The values between 0.2 and 0.4 are considered a fair agreement. Given the subjective nature of stress perception, and as many as 11 discrete points, we consider this agreement reasonably good for this task.

It should be kept in mind, however, that Fleiss' kappa considers the discrete rating as independent of each other and penalizes a one-step difference between two annotators (e.g. 2–3) in the same way as a seven-point difference (2–9). To capture the distance among annotators, we also calculated variance and standard deviation for each sentence; average of values for all sentences is shown in Table 2.

Original annotations			Normalized annotations		
Fleiss	Variance	stdev	Fleiss	Variance	stdev
0.31	1.22	9.71	0.35	0.63	7.84

Tab. 2. Evaluation of inter-annotator agreement

5.3 Intended vs. perceived level of stress

Figure 4 shows how the intended levels of stress produced by the actors corresponds to the levels of stress perceived by the annotators. We plotted average stress ratings for three levels of stress and three categories of stressful situations from Table 1 in section 2.3. The figure provides several initial observations. First, the ratings show that the actors were consistently successful in separating the three levels of stress. Second, there is a difference between sentences in completely neutral situations and the fourth bar of Level 1 and the other three bars, i.e., those acted out in a neutral way but including stress semantically. This difference may stem either from the effect of text semantics on the actors, the annotators, or both. Third, the situations grouped under psycho-social stress are perceived/produced as less stressful than the situations in the other two groups consistently at all three stress levels. We may speculate that the nature of these situations (at home with kids or at work with a colleague) elicits lower stress levels either due to the less severe stressors, or certain amount of control over the situation compared to the other two groups involving less control and greater severity.

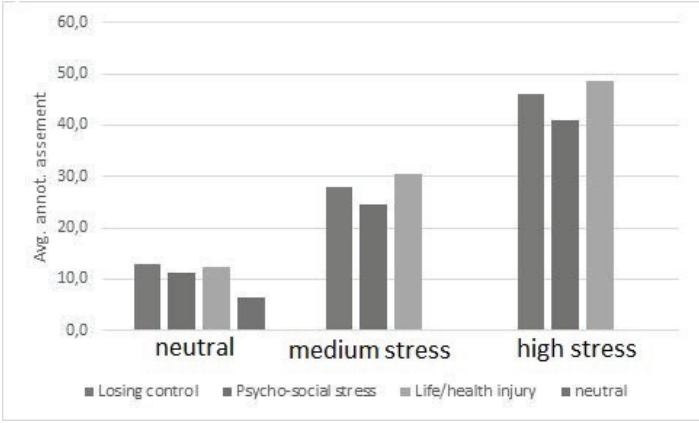


Fig. 4. Comparison of the average evaluation of annotators in relation to the played level of stress and the classification of situations into groups

6 DISCUSSION AND FUTURE WORK

The sampling of both the actors and the annotators provides richness and variability in that each utterance from the corpus is produced by multiple speakers and its stress level is assessed by multiple annotators. Thus, the information about the intended level of stress in speech production and the associated perceived level of stress for each utterance of StressDat provide the basis for developing the statistical models predicting the level of stress in speech.

The complete database will contain 30 speakers, and will be accessible for research purposes.

ACKNOWLEDGMENTS

This work is supported by European Union's Horizon 2020 research and innovation programme under grant agreement number 832969, project SATIE "Security of Air Transport Infrastructure of Europe". This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein. For more information on the SATIE project see: http://satie-h2020.eu/. This work was also funded by the Slovak Scientific Grant Agency VEGA, grant number 2/0161/18.

References

[1] Yogesh, C. K. et al. (2017). Bispectral features and mean shift clustering for stress and emotion recognition from natural speech. In Computers & Electrical Engineering, 62, pages 676–691.

[2] Robinson, C., and Nicolas, R. A. (2019). Sequence-to-sequence modelling of f0 for speech emotion conversion. In ICASSP 2019, pages 6830–6834.
[3] Rusko, M., Trnka, M., Darjaa, S., Stelkens-Kobsch, T., and Finke, M., (2018). Weaknesses of voice biometrics – sensitivity of speaker verification to emotional arousal. In ICSV25: 25th International Congress on Sound and Vibration. Hiroshima, Japan, pages 1–8.
[4] Alimuradov, A. K. et al. (2020). Development of Natural Emotional Speech Database for Training Automatic Recognition Systems of Stressful Emotions in Human-Robot Interaction. In 4th Scientific School on Dynamics of Complex Networks and their Application in Intellectual Robotics (DCNAIR), pages 11–16.
[5] Busso, C. et al. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. In Language resources and evaluation, pages 335–359.
[6] Hansen, J. H. et al. (1997). Getting started with SUSAS: a speech under simulated and actual stress database. In Eurospeech, 97(1), pages 1743–1746.
[7] Burkhardt, F. et al. (2005). A database of German emotional speech. In Ninth European Conference on Speech Communication and Technology.
[8] Campbell, N. (2000). Databases of emotional speech. In ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.
[9] Rusko, M., Darjaa, S., Trnka, M., Sabo, R., and Ritomský, M. (2015). Expressive Speech Synthesis for Critical Situations. COMPUTING AND INFORMATICS, 33(6), pages 1312–1332.
[10] Enos F., and Hirschberg J. (2006). A framework for eliciting emotional speech: Capitalizing on the actors process. In First International Workshop on Emotion: Corpora for Research on Emotion and Affect LREC 2006, Genoa, Italy, pages 6–10.
[11] Wolpe, J. (1969). The Practice of Behavior Therapy. Pergamon Press, 314 p.
[12] Accessible at: https://ccp.net.au/suds-thermometer/.
[13] SATIE project: "D4.2 – Traffic Management Intrusion and Compliance System", Status: submitted.
[14] Accessible at: https://en.wikipedia.org/wiki/Standard_score.
[15] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5), pages 378–382.

sciendo
DOI 10.2478/jazcas-2021-0054

LINGUISTIC ANNOTATION OF TRANSLATED CHINESE TEXTS: COORDINATING THEORY, ALGORITHMS AND DATA

KIRILL I. SEMENOV[1] – ARMINE K. TITIZIAN[2]
– ALEKSANDRA O. PISKUNOVA[2] – YULIA O. KOROTKOVA[2]
– ALENA D. TSVETKOVA[2] – ELENA A. VOLF[2]
– ALEXANDRA S. KONOVALOVA[2] – YULIA N. KUZNETSOVA[3]
[1] Charles University, Prague, Czech Republic
[2] National Research University "Higher School of Economics", Moscow, Russia
[3] Lomonosov Moscow State University, Moscow, Russia

SEMENOV, Kirill I. – TITIZIAN, Armine K. – PISKUNOVA, Aleksandra O. – KOROTKOVA, Yulia O. – TSVETKOVA, Alena D. – VOLF, Elena A. – KONOVALOVA, Alexandra S. – KUZNETSOVA, Yulia N.: Linguistic annotation of translated Chinese texts: Coordinating theory, algorithms and data. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 590 – 602.

Abstract: The article tackles the problems of linguistic annotation in the Chinese texts presented in the Ruzhcorp – Russian-Chinese Parallel Corpus of RNC, and the ways to solve them. Particular attention is paid to the processing of Russian loanwords. On the one hand, we present the theoretical comparison of the widespread standards of Chinese text processing. On the other hand, we describe our experiments in three fields: word segmentation, grapheme-to-phoneme conversion, and PoS-tagging, on the specific corpus data that contains many transliterations and loanwords. As a result, we propose the preprocessing pipeline of the Chinese texts, that will be implemented in Ruzhcorp.

Keywords: Mandarin, Russian, parallel corpus, Chinese word segmentation (CWS), grapheme-to-phoneme conversion (G2P), PoS-tagging, code-switching detection

1 INTRODUCTION

Linguistic annotation is one of the key concepts for current corpus linguistics. Chinese has its unique aspects of linguistic annotation, namely: absence of word segmentation conventions; Chinese characters (with a high degree of homophony and homography); significant distinctions in the morphosyntactic system between the Chinese and the European languages. All the above-mentioned problems are compounded if a sentence contains loanwords, as their phonological, morphological, and orthographic features usually contradict the standard parameters of Chinese words.

The problem of proper annotation of Chinese texts that contain loanwords and transliterations became crucial for the project of the Russian-Chinese parallel corpus (hereinafter – Ruzhcorp; [1]) – a project within the Russian National Corpus. The collection of texts in Ruzhcorp comprises 1070 documents, and the total number of tokens (Russian and Chinese) is more than 3.5 million. The majority of the texts belongs to the fiction domain (81%), and news articles (11%).

Until recently, the linguistic annotation of Ruzhcorp has been inappropriate. The Chinese word segmentation algorithm (henceforth – CWS) was based on a variant of simple greedy search over a pre-loaded dictionary. This caused problems with the detection of Russian loanwords, as they are usually absent in the dictionaries. The algorithm of pinyin (official romanisation in PRC) attribution assigned all the possible readings to each of the characters. Finally, there was no morphosyntactic annotation (hereinafter – PoS-tagging) for Chinese texts.

Our research was aimed to create a proper pipeline of linguistic annotation of the Chinese texts for Ruzhcorp, that would: i) be consistent regarding the linguistic theory; ii) show appropriate results for the original Chinese texts; iii) show appropriate results for detection of Russian transliterations and loanwords in the Chinese texts. Speaking about the necessary layers of annotation, the pipeline should include CWS, PoS-tagging, and pinyin annotation (hereinafter – G2P from "grapheme-to-phoneme"). Within this article, we are going to provide an overview of aspects i and iii, as, relating to aspect ii, we are relying on the analyses carried out by the research community. In Part 2, we present the theoretical comparison of the CWS and PoS-tagging standards for Chinese. In Part 3, we describe the set of experiments in CWS, PoS-tagging and G2P on the Chinese texts of Ruzhcorp. In Part 4, we propose the final model for Chinese linguistic annotation for Ruzhcorp, based on our theoretical and empirical comparisons.

2 THEORETICAL COMPARISON OF THE STANDARDS OF CHINESE LINGUISTIC ANNOTATION

2.1 Chinese word segmentation

The concept of "word" in Chinese is a challenging issue. Firstly, there are no spaces in Chinese, and secondly, a character, not a word, was traditionally considered a linguistic unit. But with the growing necessity of tokenization for different NLP tasks, several segmentation standards were developed – [2]. Every standard tends to focus on one of the language levels: morphosyntax, semantics or lexicology. The comparative table of the standards is shown in Table 1.

Standard (Abbreviation)	Basic principle	Description
GB T 13715-1992	lexicography, semantics	The oldest standard, implemented in mainland China in 1993. The standard lacks theoretical foundation and seems too arbitrary.
Peking University standard (PKU)	lexicography, semantics	Based on GB T 13715-1992 with some rules redefined. Segmentation units are determined by lexical semantics and lexical combinability.

Standard (Abbreviation)	Basic principle	Description
CNS 14366 (hereinafter – CNS)	morphology, syntax, semantics	Taiwanese national standard, implemented in 1999 by Academia Sinica. The standard operates with various linguistic concepts, such as morpheme, affix, dependent word and so on, and its rules are more consistent.
Penn Chinese Treebank standard (CTB)	syntax	Based on X-bar syntax theory, with every constituent that can take X^0 position considered as a word. The standard was created specifically for Chinese treebank, and is compatible with Universal dependencies tagset.
Microsoft Research Asia (MSR)	morphology, syntax, semantics	Developed its word taxonomy: lexical word, morphologically derived word, factoid, new word and named entity, all of them are processed in different way. This standard is not self-sufficient and oriented towards compatibility with others: PKU, CNS and CTB.
Vocabulary standards	lexicography	Not holistic standards, because the main rule for them is to consider as a word every unit that is found in a vocabulary.

Tab. 1. Comparison of CWS standards

As we can see, CNS and CTB standards seem to be more systematic and have a strong theoretical background, which makes them both preferable standards.

2.2 PoS-tagging

Tagsets for automatic annotation of Chinese also vary in criteria for parts of speech distinction and number of categories.

One of the most widely used tagsets is Peking University morphosyntax-based standard (PKU) and its modifications. It includes 26 basic categories and up to 46 subcategories, including denoting semantic and morphological classes within basic categories.

The ICTCLAS tagset was made by the Institute of Computer Science, Chinese Academy of Sciences. This is one of the few standards for Chinese that proposes a hierarchical model of morphosyntactic tags with three levels, where the first one denotes parts of speech, and the two latter denote other categories (primarily semantic ones). This standard is one of the most numerous, with more than 90 different tags.

Chinese National Standard (CNS) has a highly detailed list of about 150 tags. Although the main criterion for selection is morphosyntactic properties, the categories highly depend on semantics as well. Although the explanatory power of this standard is high, due to the number of tags, this standard is difficult to be implemented by automatic taggers.

Universal Dependencies (UD), a syntax-based tagset, offers only 15 to 17 clear categories, which makes it convenient for cross-language annotation, but appears to be less distinctive for Chinese than it should be.

Finally, Penn Chinese Treebank (CTB) 3.0, which was a prototype for Chinese UD, is based on the principle of syntactic distribution and has 33 tags. The moderate number of tags and the principles of their attribution that can be modelled through programming means make CTB the most applicable PoS standard.

3 EXPERIMENTS WITH LINGUISTIC ANNOTATION ON RUZHCORP DATA

3.1 Data

Ruzhcorp data have substantial differences from the "standard" Chinese texts, as they contain phonetical borrowings and transliterations, which sum up to several thousand. The majority of these unusual tokens occur either in the texts translated from Russian or in the texts that describe Russian realities. Most of these tokens constitute transliterations of Russian proper names (toponyms and anthroponyms), thus, hereinafter we will focus only on the phonetical transliterations of the proper nouns and will use "loanwords" and "transliterations" as synonyms.

To evaluate the performance of the algorithms that cover features of CWS, PoS-tagging and G2P, we created the datasets on Ruzhcorp data, which share three common features:

1. Separate datasets for fiction (Russian-to-Chinese translations) and news domains (articles in Chinese media about Russia).
2. Sentences in each dataset are balanced (each document does not exceed 8–10% of the dataset) and randomized.
3. Objects in each dataset have common features (Russian and Chinese sentences) and the features specific to this dataset. These peculiarities, as well as the quantitative overview of each dataset, are presented in Table 2.

Dataset[1]	**Size (sentence pairs)**	**Features**	**Used in**	**Purpose**
BOOKS_1/NEWS_1	436/78	(automatically) Extracted Russian proper names + their (manually) extracted transliterations. Only sentences with Russian proper names.	CWS, PoS-tagging	evaluation
BOOKS_2/NEWS_2	688/158		CWS, code-switching	fine-tuning
BOOKS_3/NEWS_3	>800/>400		CWS (future)	fine-tuning

[1] The prefix BOOKS means the data are taken from fiction literature, NEWS – from the news articles. The "size" column values are separated by slash for BOOKS_x and NEWS_x datasets, respectively.

Dataset[1]	Size (sentence pairs)	Features	Used in	Purpose
BOOKS_G2P/ NEWS_G2P	650/700	Manual pinyin annotation of the whole sentences.	G2P	evaluation

Tab. 2. Description of the datasets

3.2 Experiments in CWS

3.2.1 Comparison of the best performing CWS algorithms without fine-tuning

Our first task was to evaluate the performance of different CWS algorithms regarding the identification of transliteration boundaries in the sentences. Firstly, we have tested the following algorithms that are widely used for the CWS task and show high quality on default Chinese texts.

Algorithms	Architecture	CWS standards
Ckiptagger [3]	neural network: bidirectional LSTM and multi-head attention layers	CNS
Stanza [4]		CTB
SpaCy [5]		CTB
Pkuseg [6]	neural network: adaptive online gradient descent	PKU
FastHan [7]	neural network: BERT	PKU, CNS, CTB, MSR (different pretrained variants)
NLPIR [8]	dictionary-based method followed by a k-shortest path routing	dictionary
LTP [9]	neural network: ELECTRA	PKU
UDPipe [10]	neural network: bidirectional GRU	CTB

Tab. 3. Overview of the considered CWS algorithms

To compare the algorithms, we used two datasets – BOOKS_1 and NEWS_1. We applied all the above-mentioned algorithms to the datasets and calculated three metrics for each algorithm: recall, F-score, and our metric (hereinafter – "our") that penalizes models for both overtokenization (segmentation of one loanword into more tokens) and undertokenization (setting broader boundaries for a loanword than necessary). The original metric was designed because traditional metrics do not properly reflect the boundaries of the tokens, rather aiming at their number in a sentence. The results on the Ruzhcorp data are represented in the following figure.

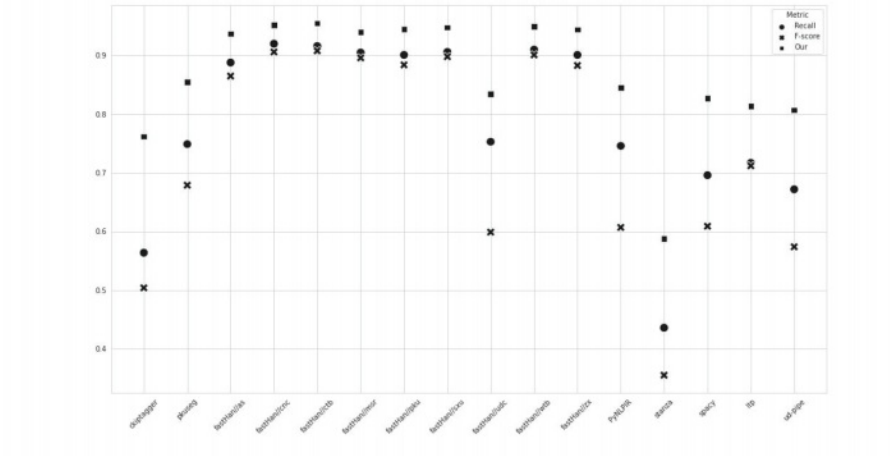


Fig. 1. Results of CWS algorithms on BOOKS_1 data

The quantitative analysis demonstrates, firstly, that despite its widespread, graph-based model, NLPIR was not as good as neural networks. Secondly, monolingual Chinese models turned out to be better than multilingual Stanza and UDPipe. Thirdly, in some cases, the performance of algorithms correlates with the CWS standards (CNS-based Ckiptagger performs worse than PKU-based PKUSeg). However, we cannot state a causal link, because fastHan, which has a CNS variant, performs as well as its PKU or CTB versions.

The qualitative analysis made us discover some inconsistencies within CWS standards and algorithm performance. For instance, the multi-word Russian personal names (e.g, first name and patronymic) are divided only by some segmenters, such as the middle dot – a special symbol "·" in the Mandarin orthography. The standards regard such clusters variously as well: PKU and CNS prescribe not to divide them, while CTB does not. We believe that it is necessary to split the multi-word transliterations by the middle dot, as this symbol is proof that the Chinese speakers are aware of the multi-word nature of these items.

Based on our analysis, we identified fastHan and its CTB-based versions (like fastHan//ctb or fastHan//wtb) as the best algorithm for our Corpus.

The detailed results of the study are represented in [11].

3.2.2 Experiment with fine-tuning FastHan algorithm

Another advantage of fastHan is a built-in fine-tuning function. Thus, we decided to test whether the fine-tuned algorithms would perform better on our data. We fine-tuned the best-performing variants of fastHan based on three main CWS standards: CNS, CTB, PKU. We used BOOKS_2 and NEWS_2 datasets. We passed datasets to CNS-based, CTB-based and PKU-based models, accordingly.

Unexpectedly, the performance of the fine-tuned algorithms after testing on BOOKS_1 and NEWS_1 slightly degraded. The table provides a comparison between models' metrics before fine-tuning and after it.

	Before fine-tuning			After fine-tuning		
Metric	**FastHan// PKU**	**FastHan// CTB**	**FastHan// CNS**	**FastHan// PKU**	**FastHan// CTB**	**FastHan// CNS**
Recall	0.9030	0.9180	0.9214	0.8997	0.9064	0.9080
F-score	0.8860	0.9097	0.9077	0.8841	0.8988	0.8902
Our	0.9488	0.9579	0.9554	0.9468	0.9493	0.9493

Tab. 4. Comparison of the FastHan algorithms before and after fine-tuning

We suggest that the main reason was the following: there was a small overlap between the set of proper names in our BOOKS_1+NEWS_1 (test) and BOOKS_2+NEWS_2 (fine-tuning) datasets, as different documents were taken. The overlap comprises only 10 words, which is less than 10% overlap in the test dataset and less than 5% – in the training dataset, thus, the model did not "learn" how to tokenize exact proper nouns in the test dataset.

Currently, we are compiling another dataset – BOOKS_3 and NEWS_3, sharing the same text sample as in a test dataset and being of bigger size, to proceed with experiments in more representative fine-tuning.

3.2.3 Experiment in code-switching detection

Another hypothesis for handling the problem of transliterations was not to fine-tune the CWS models but to use a different module that would be aimed only at code-switching detection, which, in our case, would mean the transliterated Russian proper nouns. As we approach this task, it can be treated as sequence labelling.

To do this, we ascribed labels to the transliterations in BOOKS_2 dataset and trained the LSTM and CRF layers of fastHan algorithm on it. The NEWS_2 dataset was used to check the performance on out-of-domain data.

The results of the experiment are decent, as the table below represents, however, we do not consider them reasonable to add the gained increase to the main pipeline because of lack of training data. Moreover, the performance on OOD data is worse than the original fastHan, thus we conclude that this technique to adjust the quality is needless for our task.

	Fine-Tuning Data (BOOKS_2)			**Test on out-of-domain data (NEWS_2)**		
Metric/Model	FastHan//CNS	FastHan//CTB	FastHan//PKU	FastHan//CNS	FastHan//CTB	FastHan//PKU
Recall	0.8990	0.9296	0.9292	0.7861	0.8181	0.8094
F-score	0.8824	0.9189	0.9226	0.7712	0.8054	0.7882

Tab. 5. Results of code-switching detection experiment

3.3 Experiments in G2P

The main problem of the G2P task for Chinese is that many Chinese characters have multiple phonetic representations depending on the word or syntactic position, so disambiguation of the readings for each character appears to be the key challenge for the task. In general terms, Chinese G2P annotation includes the following steps: word segmentation (and possibly PoS-tagging), obtaining all possible phonetic values for a token, and applying a set of heuristics to choose the most relevant transcription. Therefore, the quality of the pinyin annotation depends on the quality of the previous part(s) of a pipeline, CWS, and PoS-tagging.

In this study, we tested the following G2P algorithms for pinyin annotation. G2pC [12] is based on recurrent neural networks. G2pM [13] is a package with bidirectional LSTM architecture. The Xpinyin [14] model is based on stochastic decision lists using frequencies of pinyin. Pypinyin [15] library uses n-gram statistics and has an in-built collocation dictionary. The G2pC model is the only one to use an external application for CWS and PoS-tagging. Thus, we used the G2pC model with different tools for CWS: PKUSeg, a default model, fastHan and UDPipe.

For the test, we used two manually annotated datasets, BOOKS_G2P and NEWS_G2P, consisting of 1350 annotated sentences. For each character, a pinyin annotation was ascribed. Table 6 presents accuracy scores on the test dataset for each model.

Model	Performance (Accuracy)
G2pC (PKUSeg)	0.7347
G2pC (FastHan)	0.7304
G2pC (UDPipe)	0.7239
G2pM	0.5607
Xpinyin	0.5457
Pypinyin	0.5459

Tab. 6. Comparison of the phonetic annotation results

The best model is G2pC with PKUSeg word segmenter. PKUSeg is pre-trained on several datasets of different domains (medicine, art, etc.) which may help it perform on new data better than other models which are mainly trained on news texts. However, G2pC with fastHan word segmentation shows almost the same performance as the default CWS model.

The detailed results are represented in [16].

3.4 Experiments in PoS-tagging

3.4.1 Comparison of the best performing Chinese PoS-taggers

Regarding PoS-taggers, our first interest was to compare their performance on transliterated toponyms and anthroponyms specifically. For the first PoS-tagging

experiment, we examined a group of algorithms represented in the Table below. We can see that almost every tool uses a different tagset, which were compared in 2.2. Regarding the problem of loanwords, we distinguished three groups of tags: for anthroponyms; for toponyms; for more common classes or other lexical classes of the proper names (for example, common nouns or all nouns).

Tool	PoS Tagset	Anthroponyms	Toponyms	More common and related classes
Ckiptagger [3]	Chinese national standard (CNS)	Nb	Nc	Na
PKUSeg [6]	Peking university (PKU)	nr	ns	n, nz
FastHan [7]	Penn Chinese Treebank (CTB)	NR	NR	NN
PyNLPIR [17]	PKU (modified)	nrf	nsf	n, nr, ns, nt, nz
Stanza [4]	Universal Dependencies + CTB (UPOS)	PROPN	PROPN	NOUN
SpaCy [5]	UPOS	PROPN	PROPN	NOUN
LTP [9]	PKU (modified)	nh	ns	n, ni, nz

Tab. 7. Comparison of Chinese PoS-tags that can be classified as borrowings

To compare the algorithms, we used BOOKS_1 and NEWS_1 datasets by taking the sentences, splitting them with CWS algorithms, and applying PoS-taggers. After that, we evaluated the PoS-tags of transliterations. We divided the ascribed PoS-tags into three groups: absolutely correct (when the tagger matches both the part of speech and the semantic class of the word), approximate match, when the tagger chose a morphosyntactically correct annotation but did not ascribe the exact lexical class (for instance, an anthroponym was marked as a toponym or a common noun), and all other cases that are error. The algorithms were evaluated by the F-score metric (Fig. 2).

According to the results, the best tool is fastHan, which has almost 100% correctness. The main errors of all algorithms occurred due to incorrect word segmentation (thus we did not analyse them precisely). Speaking about the mistakes among the correctly segmented words, a notable inaccuracy was marking anthroponyms as toponyms and vice versa. The possible explanation is that such words end with the morphemes that are usually used as semantic markers of the proper names from the opposite groups, so the tagger could decipher them as a generic element (see Conclusions) rather than the last character of the transliteration. The detailed results are represented in [18].

3.4.2 Experiments in parallel PoS-tagging

The method of parallel PoS-tagging is gaining popularity for the multilingual data: among two languages in the parallel corpus, the well-studied standard language for which the task of PoS-tagging is relatively well solved is used as an additional

sequence of tags for labelling the under-resourced language. This approach was used either for low-resourced languages or for languages with grammar that differs significantly from European languages.

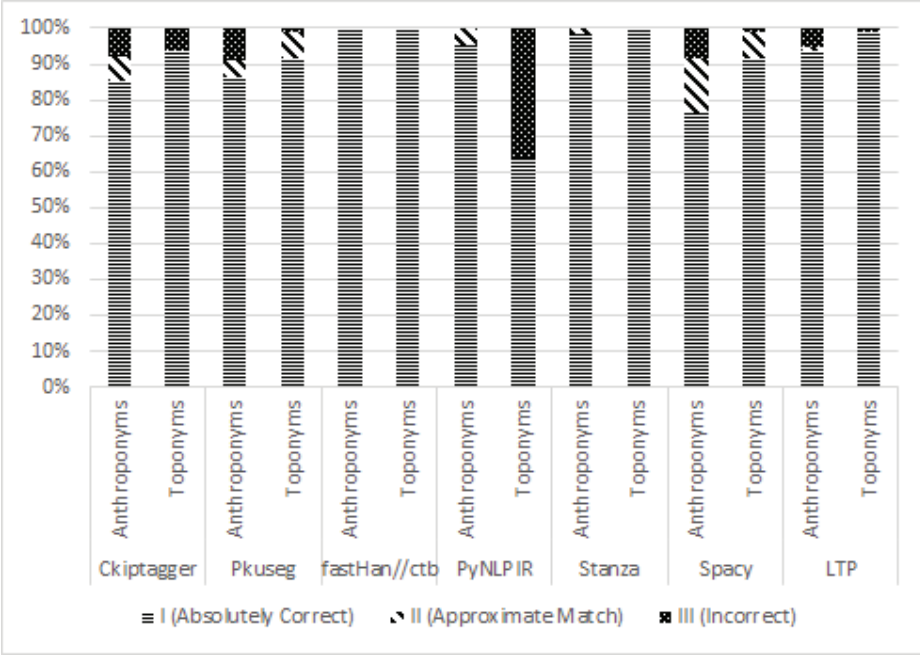


Fig. 2. PoS-taggers' results on our dataset

We implemented this approach to Ruzhcorp data. For that work, we used BOOKS_2 and NEWS_2 datasets, with manual word-to-word alignment. The data were divided at a ratio of 7 to 3 into the training and the test sets. Russian sentences were parsed by a morphological annotator Pymorphy2 and the Chinese were processed by fastHan. Then the two sequences of PoS-tags were given as two inputs for a BiLSTM-based neural network. The accuracy on the test set reached a slightly better score of 0.98 compared to 0.97 produced by a default fastHan model. We consider that an interesting source for further research, however, this approach appears to be excessive in production as it requires word-to-word alignment of data and cannot handle raw text.

The detailed results of the experiment are described in [19].

4 THE FINAL MODULE FOR LINGUISTIC ANNOTATION OF THE CHINESE TEXTS IN RUZHCORP

After reviewing all the results, we merged all the modules into one algorithm of Chinese text processing. Our algorithm consists of CWS, PoS-tagging, and G2P

functions and a module with custom rules that split the multi-word transliterations by the middle dot (see 3.2.1). Each module is applied sequentially: the CWS, PoS-tagging, G2P and custom rules are applied one after another and take the outputs of the previous module as inputs.

While working on this algorithm, we had to decide on which standard it should be based and which tools we should use for each task. Our decision is based not only on the idea of using the best standards and tools for each task, but also on the idea of making all the tools work in harmony in our algorithm. In case of CWS and PoS-tagging tasks, there is no problem as both of these modules perform best results using fastHan based on CTB standard. This standard is considered the best for processing Chinese texts because, unlike other standards, it is centred on syntactic structure, which is more relevant to Chinese than morphological and lexico-semantic features. However, considering the G2P task, it is a little more difficult because the best tool for it is G2pC, which uses PKUSeg as a word segmenter and a PoS-tagger. Nevertheless, we decided to implement CTB-based fastHan into G2pC although this implementation performed negligibly worse performance than the default G2pC as was shown in Section 3.3.

G2pC, unlike other tools, takes into consideration CWS and PoS-tagging annotation and uses this information to solve the ambiguity problem, which explains its good performance. All other algorithms take only CWS as input, which logically lowers their results. This allows us to conclude that it is more rewarding to create a "sequential" structure of CWS, PoS-tagging, and G2P modules (each module takes as input the output of the previous module) rather than a "parallel" structure (PoS-tagging and G2P modules take only CWS results independently).

The code-switching detection was not included because it works worse than algorithms for the CWS task on their own. Parallel PoS-tagging showed better results than monolingual PoS-taggers, but it cannot be used for the annotation in our corpus, at least for now, as it requires Russian sentences with a deeper manual markup.

The code is available through this link: https://github.com/ruzhcorp/ruzhcorp_chinese_annotation.

5 CONCLUSIONS AND PERSPECTIVES

The paper presents a comparative analysis of the current situation in Chinese word segmentation, PoS-tagging, and automatic transliteration from both theoretical and experimental sides by using Ruzhcorp data. In terms of theory, the frameworks that fit the Russian-Chinese parallel corpus most are the syntax-based standards of both CWS and PoS-tagging (such as CTB or UD) and that the best G2P predictions are made with the use of information about tokenization and PoS-tags. From the technical perspective, the best algorithms are, firstly, based on the modern neural

architectures (namely BERT, ELECTRA and RNN). Secondly, for Chinese-specific tasks like CWS, monolingual algorithms perform better than multilingual ones.

As a result of the set of experiments, we propose an algorithm that includes all three aspects of the Chinese linguistic annotation, and that features both neural and rule-based patterns. To date, all the texts in Ruzhcorp have been re-annotated with this algorithm and are available at the webpage https://ruzhcorp.github.io/.

There are areas of future research in that field. Firstly, our observations show inconsistencies in the detection of the so-called generic elements in Chinese: after a proper noun, a "generic" noun is used in order to denote the type of objects the name refers to. The CWS standards treat this phenomenon in significantly different ways, taking into account phonotactic (length of the generic element) or semantic features. Thus, we find it necessary to provide a specification of the CWS standard for Ruzhcorp, which will include a more consistent approach to generic elements. The second research area is deepening the experiments on parallel linguistic annotation. On the one hand, this can be conducted for scientific purposes, such as parallel PoS-tagging, on the other hand, this is a valuable help for the task of word-to-word alignment, which is rather aimed at corpus-aided language learning.

ACKNOWLEDGEMENTS

Our work is supported by the Commission for the Support of Educational Initiatives of the Faculty of Humanities of HSE University (project title – "Language-Specific Markup of Chinese Texts in the Russian-Chinese Parallel Corpus of RNC").

References

[1] Semenov, K. I., Kuznetsova, Y. N., and Durneva, S. P. (2020). Russian-Chinese parallel corpus of RNC: Problems and perspectives. Proceedings of the 10th International Conference "Russia and China: History and Perspectives for Cooperation", pages 633–640.

[2] Emerson, T. (2005). The Second International Chinese Word Segmentation Bakeoff. Accessible at: http://sighan.cs.uchicago.edu/bakeoff2005/.

[3] Li, P.-H., and Ma, W.-Y. (2019). CkipTagger. Accessible at: https://github.com/ckiplab/ckiptagger.

[4] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. Association for Computational Linguistics (ACL) System Demonstrations. Accessible at: https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

[5] Honnibal, M., and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Accessible at: https://spacy.io/.

[6] Luo, R., Xu, J., Zhang, Y., Ren, X., and Sun, X. (2019). PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation. Accessible at: http://arxiv.org/abs/1906.11455.

[7] Geng, Z., Yan, H., Qiu, X., and Huang, X. (2020). fastHan: A BERT-based Joint Many-Task Toolkit for Chinese NLP. Accessible at: http://arxiv.org/abs/2009.08633.
[8] Zhang, H., and Shang, J. (2019). NLPIR-Parser: An intelligent semantic analysis toolkit for big data. Corpus Linguistics, 6(1), pages 87–104.
[9] Che, W., Feng, Y., Qin, L., and Liu, T. (2021). N-LTP: A Open-source Neural Chinese Language Technology Platform with Pretrained Models. Accessible at: http://arxiv.org/abs/2009.11616.
[10] Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 197–207. Accessible at: https://doi.org/10.18653/v1/K18-2020.
[11] Semenov, K. I., Korotkova, Y. O., Volf, E. A., and Konovalova, A. S. (2021). Automatic Annotation of the Chinese Texts that Contain Loanwords: Word Segmentation, Transcription, PoS-tagging. DIALOG-2021: 27th International Conference on Computational Linguistics and Intellectual Technologies, Supplementary volume, pages 1081–1095. Accessible at: http://www.dialog-21.ru/media/5420/_-dialog2021supvol.pdf.
[12] Cai, Z., Yang, Y., Zhang, C., Qin, X., and Li, M. (2019). Polyphone Disambiguation for Mandarin Chinese Using Conditional Neural Network with Multi-level Embedding Features. Accessible at: https://arxiv.org/abs/1907.01749.
[13] Park, K., and Lee, S. (2020). g2pM: A Neural Grapheme-to-Phoneme Conversion Package for Mandarin Chinese Based on a New Open Benchmark Dataset. Accessible at: http://arxiv.org/abs/2004.03136.
[14] Luo, E. (2020). Xpinyin. Accessible at: https://github.com/lxneng/xpinyin.
[15] Huang, H. (2020). pypinyin. Accessible at: https://github.com/mozillazg/python-pinyin.
[16] Konovalova, A. S., and Tsvetkova, A. D. (2021). Comparative analysis of grapheme-to-phoneme models for the Russian-Chinese parallel corpus. Program book of Buckeye East Asian Linguistics Forum 4, pages 28–30. Accessible at: https://cpb-us-w2.wpmucdn.com/u.osu.edu/dist/6/3609/files/2021/03/BEALF-4_Program_Book_2021-3-5.pdf.
[17] Roten, T. S. (2018). PyNLPIR PoS tagset. Accessible at: https://pynlpir.readthedocs.io/en/latest/api.html.
[18] Semenov, K. I., Korotkova, Y. O., and Volf, E. A. (2021). Automatic Annotation of the Russian Loanwords in Chinese Texts: Issues in Word Segmentation and PoS-tagging. Proceedings of Corpora 2021 International Conference. 14 pages [in press].
[19] Konovalova, A. S. (2021). Automatic POS-tagging for Chinese Using Parallel Data [BA thesis]. Higher School of Economics. 82 pages.

sciendo
DOI 10.2478/jazcas-2021-0055

A ROBUST APPROACH TO VARIATION IN CARPATHIAN RUSYN: RESAMPLING-BASED METHODS FOR SMALL DATA SETS

MOULAY ZAIDAN LAHJOUJI-SEPPÄLÄ – ACHIM RABUS
Slavonic Institute of the Albert-Ludwig-University of Freiburg, Freiburg, Germany

LAHJOUJI-SEPPÄLÄ, Moulay Zaidan – RABUS, Achim: A robust approach to variation in Carpathian Rusyn: Resampling-based methods for small data sets. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 603 – 617.

Abstract: Quantitative, corpus based research on spontaneous spoken Carpathian Rusyn language can cause several data-related problems: Speakers are using ambivalent forms in different quantities, resulting in a biased data set – while a stricter data-cleaning process would lead to a large scale data loss. On top of that, polytomous categorical dependent variables are hard to analyze due to methodological limitations. This paper provides several approaches to face unbalanced and biased data sets containing variation of conjugational forms of the verb *maty* 'to have' and *(po-)znaty* 'to know' in Carpathian Rusyn language. Using resampling based methods like Cross-Validation, Bootstrapping and Random Forests, we provide a strategy for circumventing possible methodological pitfalls and gaining the most information from our precious data, without trying to p-hack the results. Calculating the predictive power of several sociolinguistic factors on linguistic variation, we can make valid statements about the (sociolinguistic) status of Rusyn and the stability of the old dialect continuum of Rusyn varieties.

Keywords: oral corpora, border effects, language variation, spoken language corpus, robust statistics, Carpathian Rusyn

1 INTRODUCTION

As the size of empirical data and the number of bigger corpora rose as steadily as processing power of computers, complex statistical methods have obtained more and more approval in the field of linguistics. This trend also applies to dialectology and sociolinguistics – subfields with a greater focus on variation in spoken language. Compared to statistical methods applied to written language data, spoken language data can evoke several data related problems. As oral corpora are often unequally smaller than written corpora, results and the application of statistical tests have to be treated with special caution. Working with smaller data sets, outliers as well as autocorrelations between independent variables can pose the risk of causing a higher effect on the result of estimations or elaborated statistical tests than in larger, balanced data sets.

In this paper we discuss statistical methods from a sociolinguistic point of view. By analyzing a specific case of linguistic variation in Carpathian Rusyn, we

problematize the use of statistical methods by taking the rather complicated nature of spoken-language based data into account. We propose to analyze small and unevenly distributed datasets with resampling-based and robust methods, rather than reducing the complexity of the analysis or the data set for the sake of high significance levels. The aim is to avoid false positive or negative results by assessing statistics based on estimations, rather than absolute values.

The methods discuss are applied to verbal inflection in Carpathian Rusyn. The verbs *maty*, *znaty*, *poznaty*$_{3Ps.Sg.Pres.}$: 'to have' and 'to know' are analyzed with respect to the sociolinguistic embedding of the variation within the states Carpathian Rusyn is spoken, i.e.e., Poland, Slovakia, and Ukraine. The aim is to analyze which sociolinguistic factors with high influence on the outcome of the variation can be detected and whether so called "border effects" ([1], [2]) can be observed. The first section is dedicated to giving a short overview of the specific situation of Carpathian Rusyn, the background of the dataset, and the motivation of the analysis.

In the second section, the resampling methods cross validation and bootstrapping are applied to a multinomial logistic regression model, resulting in robust estimations of regression coefficients.

In the third section, we approach the variable importance via categorization with the decision-tree-bases methods Random Forest and Conditional Forest.

Since categorical dependent (and independent) variables are common in (socio-) linguistics and small, unevenly distributed data samples are more the rule than the exception when analyzing minority language data, our approaches are applicable beyond the Rusyn test case. For analyses, the open source software R-studio [3] is used.[1]

2 LINGUISTIC DATA AND METHODOLOGY

2.1 Variation in Carpathian Rusyn

Rusyn is a Slavic minority language mainly spoken in the Carpathian area, with the highest population of speakers in Transcarpathian Ukraine, Eastern Slovakia and Poland. Within the continuum of Northern Slavic languages, Rusyn is located right on the border between East- and West Slavic.

While Ukrainian is the linguistically closest language to the Rusyn varieties, their linguistic status and the national recognition of Rusyns as minorities is disputed. Some scholars claim that the Rusyn varieties are to be considered dialects of the Ukrainian language [4], others argue in favor of a separate linguistic and cultural identity of the speakers of Rusyn ([5], [6]). From a structural viewpoint, there are certain similarities with Ukrainian, e.g., with respect to common sound changes on the one hand, e.g., East Slavic polnoglasie, such as in *molodŷj* 'young' or the

[1] The R-script used for this work can be found via: https://bwsyncandshare.kit.edu/s/bGyJiGfHYkZHBa2 (please download the .html file and open with browser).

rendering of Common Slavic jat' as /i/ such as in *bilŷj* 'white'. On the other hand, certain properties make the Carpathian Rusyn varieties similar to the adjacent West Slavic languages, i.e., Polish and Slovak (for instance the use of clitic pronouns or the past tense formation using forms of the auxiliary verb 'to be' ([7], [8]). Resulting from the ambivalent status of Rusyn within the different European states, the situation is complex and dynamic. The current state of Rusyn can be researched using the online Corpus of Spoken Rusyn.[2]

In his grammar "The Rusyn Language" Stefan M. Pugh [9] describes the Prešov standardized variety of Carpathian Rusyn, from time to time with respect to other Rusyn non-standard variations (Slovak Rusyn, Lemko and Subcarpathian Rusyn). An interesting case of verbal variation is described within the conjugation classes "E(1) A(J): Conjugation Ia" and "E(2) AJ Proper" [9, p. 117–120]: The original stem marking A(J)[3] only appears in imperatives and in the non-past tense forms. As examples the verbs *čitaty* 'to read' and *maty* 'to have' are given, where the only form including the stem mark (A)J would be $čitaty_{3Ps.Pl.Pres.}$ (*čitajut'*) and $maty_{3Ps.Pl.Pres.}$ (*majut'*). The more one progressed to the east of the Rusyn dialect continuum, the more common a full A(J) conjugation would be evident (*mam* < *maju*, *mat'* < *maje*).

However, Pugh states that the A(J) forms within conjugations of this class were limited to the 3rd person plural, except the verbs *maty*, *znaty* and *poznaty*, where the A(J) forms can also be found in 3rd person singular forms. This leads to three competing forms of *maty*, *znaty*, $poznaty_{3Ps.Sg.Pres.}$:

ma, maje, mat'; zna, znaje, znat'; pozna, poznaje, poznat'.

The dataset we analyze this variation on contains 284 utterances of the above mentioned forms, by 56 speakers. The data has been obtained via query search in the Corpus of Spoken Rusyn. Corpus results can be downloaded and imported into the software R-Studio. In this case, the data set has been manually checked and cleaned[4] before the import. Besides the language samples (also available as anonymized audio recordings), the corpus also features speaker metadata (age, gender, living place, citizenship, GPS-locations).

Another variable (dialect area) has been added manually to our dataset. This variable reflects the affiliation of the villages to isoglosses that were the result of traditional dialectological research [10], before the current state borders had been established.[5] In this way, we can compare whether the traditional dialectal areas or the current states (and their respective roofing standard languages) have a stronger

[2] Accessible via www.russinisch.uni-freiburg.de/corpus (26.08.2021).

[3] Read as vowel "a" + J.

[4] We intentionally did not remove multiple utterances of the verbs by the same speaker as long as they were not within-sentence repetitions.

[5] This only applies to Rusyn data from Eastern Slovakia and Transcarpathia. The traditional Lemko dialect constellations have been torn apart by the violent resettlements of Lemko Rusyns (Akcja Wisła).

influence on the variable of interest. Statistics can also reveal differences between older and younger speakers in the sense of an apparent time study [11].

2.2 Methodological background

In order to analyze the relative importance of certain factors that might predict the outcome variable, i.e., the realization of the verb forms, we want to compare the usage of the variants of a linguistic phenomenon between several (sociolinguistic) subgroups within our data set. To do so, the variation has to be quantified. Most commonly, frequencies (of e.g. uttered word forms or specific grammatical constructions) are calculated and further on compared between several subgroups.[6] While performing statistical tests on a small data set of spoken language, quite a few problems can occur that might affect the quality of our results. Generally, researchers have a wider range of options when dealing with numerical outcome variables. Parametric statistics allows for making profound guesses about the population based on a certain underlying distribution of data, then comparing the variance within the data set with the natural distribution to assess statistical effects. However, when dealing with categorical language data there are quite a few more methodological limitations and tripping hazards.

A common statistical test in (socio-)linguistics, which is very similar to the test we are applying to our data below, is binomial logistic regression. Binomial distributions are traditionally described with a "success" "not success" scenarios like e.g. flipping coins, where each toss is independent from the latter and the probability for each side showing when it lands is equally probable (50/50 chance). The observations, derived from a random sample taken from a population are analyzed with the underlying assumption of a binomial distribution, similar to numerical data as weight and size are assumed to be distributed normally. Deviations from the distribution within the underlying population can be explained to a certain degree by factors determined within the regression formula. Besides the fact that a study design with bivariate variables can lead to (more or less necessary) oversimplification[7] of linguistic variation, the assumption of a natural 50/50 chance between two forms can as well be a bad starting point.[8]

[6] Subgroups can be defined in many possible ways and by multiple conditions. A group does not necessarily consist of many individual speakers; it could also be defined as a set of all the utterances of individuals. In our case, subgroups could consist of e.g. all female speakers. In between factor relations can be taken into account by defining and cross testing subgroups by multiple conditions (e.g. gender, age group, living place).

[7] Simplification of the variables can be an advantage form the methodological point of view because statistics involving polytomous dependent variables are disproportionately more computationally intensive and harder to analyze. The calculation of *n* baseline models can evoke to prohibitively high level of manual work and can pose the risk of bad model fitting.

[8] In researching e.g. the use of *L1* and *L2* forms, the chance of which form could be uttered may vary between individuals and groups due to random factors like weather, sympathy, geographically ambivalent perceptions of language and other factors that are hard to grasp statistically.

The most problematic property of our data set – typical for spoken and written linguistic corpora – is that many speakers utter several ambivalent forms, ranging from one utterance up to as many as twelve utterances in very different proportions. It is impossible to exclude the speakers, as we would not only have to work with an even smaller data set, but we would also willingly ignore existing variation and therefore making our whole study obsolete. For this reason, we decided to keep as much data within our sample as possible, even if this leads to some individuals dominating the data set and even though the assumption of independence between single data points is violated.

2.3 Resampling methods I: Cross validation and non-parametric bootstrapping

Taking the threefold nature of our dependent variable and several sociolinguistic factors into account, we conduct a multinomial logistic regression analysis with the formula *verb_form ~ Variety + Area + Age + Gender*. We used the function "multinom()" from the package "nnet". This function as part of the "nnet" package has several advantages such as the usual "lme4"-alike [12] regression output content and that there is no need to reshape the data set to long format. However, it does not provide p-values or t-statistics. The significance levels in Table 1 and 2 are provided by the function stargazer() of the R-package "stargazer" [13], that has been used to print the tables in HTML-format. It is important to note that this multinomial regression works by setting a baseline category and comparing two regressions side by side automatically. In our case, the set baseline of the dependent variable is the finite verb *mat* 'has'. The multinomial regression model predicts the *logit* of the two other verb forms with respect to the baseline model. As *verb_form* consists of three categories, the formula of the basic multinomial regression translates to:

$$\ln\left(\frac{P(\text{Verb\_Form} = \text{ma})}{P(\text{Verb\_Form} = \text{mat})}\right) = b_{10} + b_{11}\,(\text{Variety=Slo}) + b_{12}\,(\text{Variety=Tra}) + b_{13}\,\text{Age} + b_{14}(\text{Gender=m}) + b_{15}(\text{Area=1}) + b_{16}(\text{Area=2}) \cdots + \epsilon$$

$$\ln\left(\frac{P(\text{Verb\_Form} = \text{mae})}{P(\text{Verb\_Form} = \text{mat})}\right) = b_{20} + b_{21}\,(\text{Variety=Slo}) + b_{22}\,(\text{Variety=Tra}) + b_{23}\,\text{Age} + b_{24}(\text{Gender=m}) + b_{25}(\text{Area=1}) + b_{26}(\text{Area=2}) \cdots + \epsilon$$

Here, *P* are the odds, b_i are the regression coefficients of the respective factors and ϵ is the error term. Baselines are also set for the factors.

Trusting this *naïve* model, some model coefficients (**Tab. 1**) seem to be (highly) significant. However, we cannot solely rely on the meaningfulness of these values (even if they seem likely), not taking the violation of assumptions and the data related bias into account.

As described in **2.2**, several speakers produced different realizations of the response variable. The multinomial model cannot consider the individuals as a random factor.[9] In other words: the assumption of independence between data points is violated.

However, one can use this naïve approach in order to find a good model fit (the combination of predictive factors with the highest explanatory power) for further processing. The quality of the model fit (meaning how much of the effect can be explained by our predictors) can be assessed by comparing the Akaike Information Criterion or testing the model and measuring the accuracy.

A common approach is to split the dataset into a training and a test set in order to test the predictive power of the model on new data. For unbalanced small data sets, this approach is problematic as it worsens the model quality as some of the multiple utterances by the same speakers might be used for training and testing at the same time, moreover we would lose a greater part of our data during this process.

Alternatively, one can assess the accuracy of the model via *K-fold Cross Validation* ([15], [16], [17]). *CV* allows to split the data into *k* subsets and then compare each subset with all the other subsets. The *CV error rate is the* average *error rate* of the aggregated subset-based regression model. Doing so, the accuracy of the model can be predicted precisely without losing valuable data. The above mentioned formula has been chosen on the basis of the best *CV* accuracy rate. A formula including interaction effects between variety and age reached approximately the same accuracy rate and is therefore mentioned in the results (**Fig. 2** and **Fig. 3**). However, due to data related bias and violation of assumptions, the accuracy of the naïve regression model is merely 63%. That means the error term of the regression formulas has a predictive power of 37%.

Correlations between independent variables can have a strong influence on the outcome of the model. As shown in **Fig. 2**, the regression model with interaction effects seems to perform better (AIC) than the basic model. Nevertheless, the estimations of the regression are unreliable. The coefficient of the factor Transcarpathian variety (VarietyTRA) has become negative, even though there is no sound reason to assume a negative effect. This behavior can be explained by confounding [18]. A correlation between age and the Transcarpathian samples leads to the effect, that with the inclusion of the interaction variable, the VarietyTra:Age has not only an effect on the dependent variable, but also on the independent variable *variety*.

Hinneburg et al. [19] problematize the analyses of small datasets with a categorical dependent variable. Among other approaches, the authors show that non-parametric bootstrap can provide robust estimations of the statistics that help to avoid false assumptions about the underlying linguistic mechanisms. Fox [20] explains the principles

[9] Multinomial Logistic Mixed-Effects Regressions could potentially account for the individual variation of utterances. However, R-packages that are able to perform Mixed-Effects Regression Models for multinomial data reliably are rare. It is possible to perform several types of Multinomial Logistic Mixed-Effects Regression with the R-package "mclogit" [14] but in our case the algorithm did not converge.

behind bootstrapping regression models in R, which is that bootstrap allows estimating the distribution of regression statistics without making a priori assumptions about the distribution within the population. Therefore, data set is resampled *n*-times and the regression is calculated for each subset of the samples:

> "The essential idea of the non-parametric bootstrap is as follows: We proceed to draw a sample of size n from among the elements of S, sampling with replacement. […] The key bootstrap analogy is therefore as follows: The population is to the sample as the sample is to the bootstrap samples" [20, p. 1–2].

In this manner, not only the bias within the dataset, but also the effects of dependencies between several observations (several utterances of the same speakers) are reduced. We bootstrapped the regression model using the "boot()" function of the R-package "boot" [21].

As shown in **Tab. 2**, the median 1000-fold[10] bootstrapped regression coefficients as well as their significance levels are in the most cases less extreme than in the naïve model (**Tab. 1**).

The bootstrap process allows checking the distribution of the bootstrapped coefficients. After 1000-fold bootstrap, most coefficients seem to be normally distributed (cf. **Fig. 1**), with some of the distributions showing rather large spikes, skewness or broadly distributed minimum/maximum values. A straightforward way to calculate confidence intervals in R is by using the boot.ci() function of the package "boot" 2021 [21] or the boot_ci() function of the package "sjstats" [22]. If the distribution of bootstrapped coefficients contains larger spikes or extreme limits, boot_ci() will provide unrealistically large confidence limits for all variables. This is due to the methods being either entirely based on t-distribution or sample quantiles and the distributions are expected to be normal (no spikes, no skewness, no extreme limits). Boot.ci() provides the possibility to calculate bias-corrected and accelerated (*BCa*) confidence intervals, that seek to take skewness and bias within the distribution of coefficients into account. *BCa-CI* provide a far more realistic picture of the bootstrapped confidence intervals. **Fig. 2** displays the 95% *BCa* confidence limits of the multinomial logistic regression (without interactions), the black dots indicating the original, non-bootstrapped coefficients. Despite the fact that some *CI* are very large, the results show a more robust and less biased estimation of the coefficients. In some cases (Variety, Area), the *CI* indicates that the factors potentially have an even larger effect on the category of the dependent variable, than the median values in **Tab. 2** suggest.

[10] Meaning that the data set has been subsampled and the statistics have been calculated 1000 times.

2.4 Resampling methods II: Random Forest

When it comes to analyzing data with categorical outcome variables in R, CART[11]-based methods [23] provide a useful alternative to logistic regression models. The *bagging*[12] approach of Random Forests [24] is similar to the aggregated bootstrapping-approaches from above, but the underlying mechanisms behind CART differs from logistic regressions. Comparing CART-based models to the multinomial regression analysis (or vice versa), another equally valid perspective can be obtained. The alternative perspective can help to create a clearer picture of the calculated statistics and can, in case of a very unbalanced data set, help to verify or falsify results. Using the R-packages "randomForest" [25] and "party" ([26], [27], [28]), a robust estimation of variable importance is assessed easily, without the need to implement manual bootstrapping to the R-script. As Random Forests are even considered to be robust against presence of in-between variable interactions [29], they provide an additional corrective to the regression analysis. Without going into too much detail, we want to address a few tripping hazards that can occur while assessing the predictive power of factors via CART-based Forests.

The principle behind *decision trees* is rather straightforward. A "tree" is grown by deciding on several occasions (nodes) which factor is the most important for splitting the data between the categories of the dependent variable. Like the aggregated bootstrapped coefficients, *Random Forests* provide a robust estimate of several parameters that indicate the predictive power of factors, by combining the predictions of *n* numbers of trees, which are again based on random subsets of the data set. The data that is left out within each of the *n*-trees is used for assessing the overall accuracy of the model (OOB (out of bag)-error rate). In contrast to the regression models, *RF* algorithms use a random set of possible factors for each of these splits. It is important to check whether it is necessary to adjust the numbers of those factors. Within the formula, which is very similar to the regression formula above, the argument "mtry" indicates the amount of factors considered for each split. If "mtry" is set high, the choices between factors are less random and pose a higher risk of bias. If "mtry" is set low, the choice between factors is smaller, which may lead to a larger OOB-error rate. The OOB-error rate for the *RF* model (ntree = 10000, mtry = 3) *verb_form ~ Variety + Area + Age + Gender* was 24.7%, meaning the accuracy of the model is 75.3%.

The variable importance can be displayed with the help of the function "varImpPlot" (**Fig. 3**, left graph). The ranking of the variable importance of our analysis proves the point of Strobl et al. [29], that the mean *decrease Gini* and *mean decrease accuracy* indexes tend to be biased towards continuous independent variables (or in other cases towards variables with many categories). As shown in

[11] Classification and Regression Tree, also known as Decision Tree.

[12] Bagging is short for bootstrap aggregating.

Fig. 2, age has no predictive power in the regression models. The reason behind this error is that numerical variables like *age* can be split into various fractions, leading to considerably more options to split the decision tree branches compared to categorical variables with a very limited amount of possible splits.

Following Strobl et al. [29], the better approach for data sets with mixed (categorical and numerical) predictors is to use Conditional Forests via the function "cforest()". Conditional Forests [28], while being more computationally intensive, perform multiple significance tests at each splitting point of the trees. These significant tests (permutation tests, conceptually similar to the cross validation technique mentioned above) take several covariates of the variables into account, performing multiple significance tests on all possible combinations of predictors and covariates in the data set, preserving possible covariance structure of e.g. *variety* and *age*. As shown on the right-hand side of **Fig. 3**, the highest ranked (and therefore most important) factor is *variety*.

2.5 Interpretation

Our analysis shows that the predominating factor determining the verb forms *maty*, *znaty*, *poznaty*$_{3Ps.Sg.Pres.}$ is the factor variety, distinguishing between Transcarpathian, Lemko or Slovak Rusyn. While the old (formerly border-transgressing dialectal Areas haven't been ranked as unimportant (Area1), it seems that, at least in most cases, variety has the strongest effect. Comparing the coefficients (*ma*, *maje* vs. *mat*) of the regression model in between the varieties, Transcarpathian has by far the most homogeneous distribution of verb forms (the dominating form *maje* is congruent to the Standard Ukrainian form). Following the hypothesis of Border Effects [2] and the model of Auer and Hinskens [1, p. 17], the different embedding of Rusyn, brought about by the respective state (i.e. Poland, Slovakia and Ukraine), leads to convergence between non-standard varieties and their respective *dachsprache* and divergence within old dialectal continua. Considering the fact, that Rusyn is acknowledged as minority language in Slovakia and Poland, the more heterogeneous use of the verb forms within these varieties, including a strong use of verb forms differing from the respective umbrella languages, might not be accidental. Whereas the codified standard of Rusyn is taught in schools in Rusyn villages in Slovakia as well as in the Institute of Rusyn Language in Culture at *Prešov University*[13], the speakers of Rusyn are tending to be more confident about their language and identity [30].

3 CONCLUSION

Making correct statistical assumptions about inferences of sociolinguistic factors in spoken language data, especially dealing with a polytomous categorical variable of interest is unequally more difficult and error-prone than when dealing with parametric/

[13] https://www.unipo.sk/cjknm/hlavne-sekcie/urjk/o-institute/ (18.03.2021).

continuous data. Meeting all assumptions of the regression models and providing a balanced, unbiased data set is theoretically possible, but practically very unlikely to achieve without a prohibitively high amount of data manipulation or oversimplification of the variables of interest. The robust statistical methods suggested in this paper provide a broader perspective on the linguistic mechanisms behind the variation in spoken language, without 1. oversimplification of the data set, 2. without restricting the regression models to a binary outcome variable or just few predictors, and 3. without p-hacking. Even though the results of robust approaches are sometimes unspectacular, reporting robust estimations will reveal realistic tendencies and often significant results, instead of p-values with an unrealistically high level of significance (**Fig. 1**). By comparing several methodological approaches such as multinomial logistic regressions and Random (or Conditional) Forests, indistinct results can be re-evaluated from different points of view. As for our specific case, several statistical methods helped to uncover the underlying sociolinguistic factors behind variation within the inflectional system of verbs in Rusyn. The modern states where Rusyn is spoken have a stronger impact variation than the historical dialect areas or sociolinguistic factors such as age and gender.

It would be desirable to conduct further statistical analysis taking random factors into account as well as special factors such as the distance of the geographical location of the living place of speakers to the center of dialect areas or state borders.

ACKNOWLEDGEMENTS

The research has been funded by the German Research Foundation *DFG* under the project Number RA 2212/2-2 *"Rusyn as a minority language across state borders: a quantitative perspective"*.

Multinom. Log. Reg.: Verb Forms ~ Factors without & with Interaction Effects				
	Dependent variable:			
	ma	maje	ma	maje
	(1)	(2)	(3)	(4)
VarietySLO	-3.476***	-3.017***	2.529	-0.947
	(0.471)	(0.421)	(2.395)	(1.393)
VarietyTRA	10.800***	14.146***	**-13.792****	18.253***
	(0.355)	(0.355)	(0.001)	(0.001)
Age	0.006	0.023	0.024	0.050*
	(0.018)	(0.018)	(0.025)	(0.026)
Genderm	-1.079	-0.125	-0.876	0.163
	(0.733)	(0.724)	(0.732)	(0.740)
Area1	8.480***	11.786***	-1.262	14.462***
	(0.437)	(0.437)	(0.905)	(0.908)

Multinom. Log. Reg.: Verb Forms ~ Factors without & with Interaction Effects				
	Dependent variable:			
	ma	maje	ma	maje
	(1)	(2)	(3)	(4)
Area2	-1.155**	-0.658	-10.001***	2.845**
	(0.471)	(0.422)	(1.818)	(1.194)
VarietySLO:Age			0.034	-0.087**
			(0.056)	(0.039)
VarietyTRA:Age			0.427***	-0.056***
			(0.008)	(0.008)
Constant (mat)	3.270***	1.477	2.261*	-0.153
	(1.091)	(1.100)	(1.292)	(1.344)
Note:	*p<0.1; **p<0.05; ***p<0.01			

Tab. 1. Result table of *naïve* Multinomial Logistic Regressions model

Bootstrap Multinom. Log. Reg. Coeff. Median Values: Verb Forms ~ Factors without & with Interaction Effects R = 1000				
	Dependent variable:			
	ma	maje	ma	maje
	(1)	(2)	(3)	(4)
VarietySLO	-3.097	-2.267	0.993	0.713
	(2.857)	(1.619)	(59.64)	(34.74)
VarietyTRA	7.538**	12.564***	**-14.579**	19.052
	(3.139)	(2.739)	(24.089)	(34.87)
Age	0.007	0.026	0.0239	0.0507
	(0.020)	(0.019)	(1.087)	(1.087)
Genderm	-1.106	-0.142	0.955	0.088
	(1.551)	(1.523)	(5.237)	(5.228)
Area1	7.08	11.95***	-2.025	17.688
	(5.597)	(4.430)	(27.216)	(48.321)
Area2	-1.844	-1.608	-10.17	1.924
	(2.462)	(1.745)	(53.36)	(16.127)
VarietySLO:Age			0.041	-0.1
			(1.162)	(1.22)
VarietyTRA:Age			0.437	-0.058
			(1.68)	(1.81)

Bootstrap Multinom. Log. Reg. Coeff. Median Values: Verb Forms ~ Factors without & with Interaction Effects R = 1000				
	Dependent variable:			
	ma	maje	ma	maje
	(1)	(2)	(3)	(4)
Constant	1.41	1.4081796	2.193	-0.287
	(2.488)	(2.452)	(23.76)	(23.759)
Note:	*p<0.1; **p<0.05; ***p<0.01			

Tab. 2. Result table of *bootstrap* Multinomial Logistic Regressions models

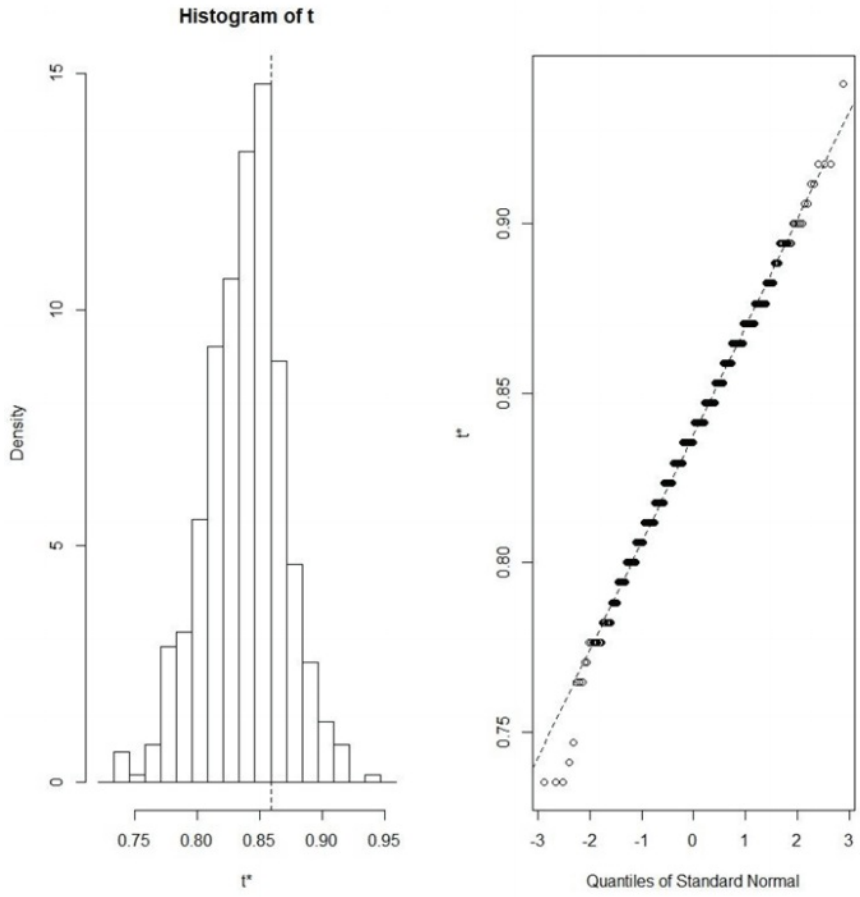


Fig. 1. Normal-like distributed bootstrap cofficients (t)

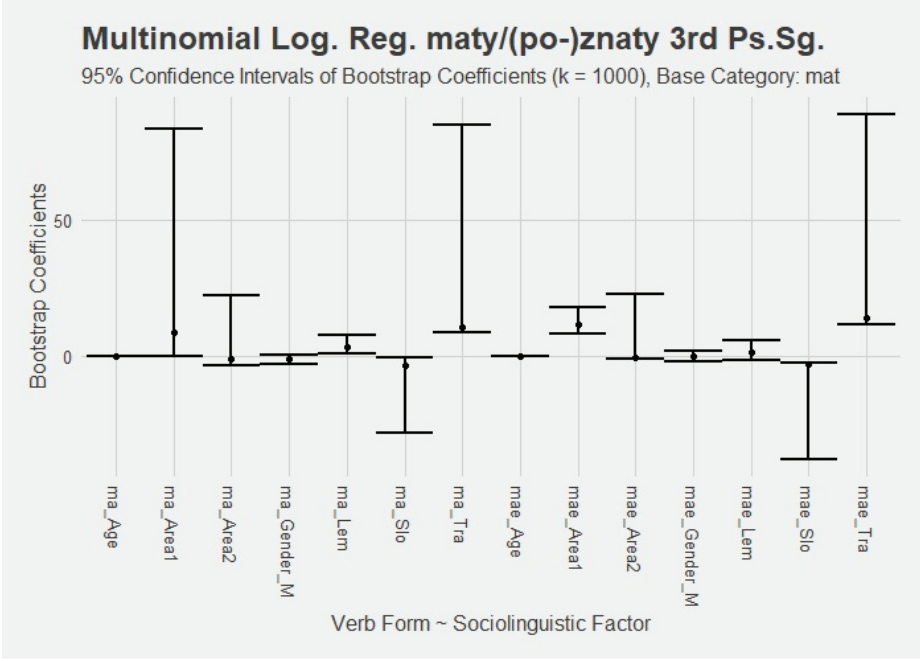


Fig. 2. Bootstrap Confidence Intervals (95%)

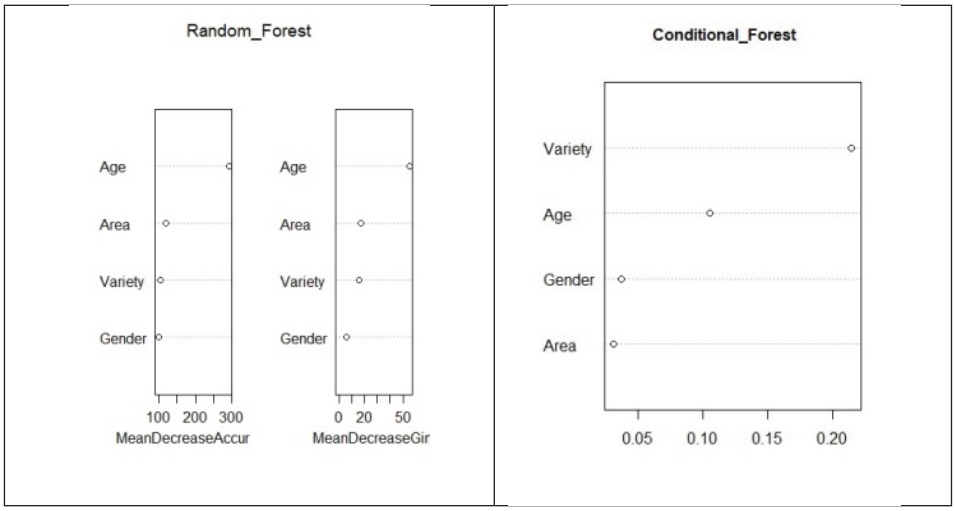


Fig. 3. Variable importance of Random Forest and Conditional Forest

References

[1] Auer, P., and Hinskens, F. (1996). Convergence and Divergence of Dialects in Europe. In Sociolinguistica (10).

[2] Woolhiser, C. (2005). Political borders and dialect divergence/convergence in Europe. P. Auer, F. Hinskens and P. Kerswill (eds.). Dialect change: Convergence and divergence in european languages. Cambridge, pages 236–262.

[3] RStudio Team. (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. Accessible at: http://www.rstudio.com/.

[4] Magocsi, P. R. (2015). With Their Backs to the Mountains: A History of Carpathian Rus' and Carpatho-Rusyns. Budapest.

[5] H. A. Skrypnyk (ed.). (2013). Ukrajinci-Rusyny: Etnolinhvistyčni ta etnokul'turni procesy v istoryčnomu rozvytku. Kyjiv.

[6] Boudovskaia, E. E. (2006). The morphology of Transcarpathian Ukrainian dialects. Los Angeles.

[7] Rabus, A. (2019). Vergangenheitsbildung in gesprochenen karpatorussinischen Varietäten: Quantitativ-statistische Perspektiven. Die Welt der Slaven 69(1), pages 15–33.

[8] Plishkova, A. (2009). Language and national identity: Rusyns south of Carpathians. Translated by Patricia A. Krafcik. With a bio-bibliographic introduction by Paul Robert Magocsi. New York (Classics of Carpatho-Rusyn scholarship, 14).

[9] Pugh, S. M. (2009). The Rusyn language: A grammar of the literary standard of Slovakia with reference to Lemko and Subcarpathian Rusyn. München (Languages of the World/ Materials, 476).

[10] Pan'kevyč, I. (1938). Ukrajins'ki hovory Pidkarpats'koji Rusy i sumežnych oblastej. Z pryložennjam 5 dialektolohičnych map. Častyna I. Zvučnja i morfolohija. Praha.

[11] Chambers, J. K. (2002). Patterns of Variation including Change. The Handbook of Language Variation and Change, pages 358–361.

[12] Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), pages 1–48. DOI 10.18637/jss. v067.i0.

[13] Hlavac, M. (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. Accessible at: https://CRAN.R-project.org/package=stargazer.

[14] Elff, M. (2020). mclogit: Multinomial Logit Models, with or without Random Effects or Overdispersion. R package version 0.8.5.1. Accessible at: https://CRAN.R-project.org/ package=mclogit.

[15] Mosteller, F., and Tukey, J. W. (1968). Data analysis, including statistics. In Handbook of Social Psychology. Addison-Wesley, Reading, MA.

[16] Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. CBSM38, SIAM, Philadelphia, Penn.

[17] Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. J. Am. Stat. Assoc., 78, pages 316–331.

[18] VanderWeele, T. J., and Shpitser, I. (2013). On the definition of a confounder. Annals of Statistics. 41(1), pages 196–220.

[19] Hinneburg, A., Mannila, H., Kaislaniemi, S., Nevalainen T., and Raumolin-Brunberg, H. (2006). How to Handle Small Samples: Bootstrap and Bayesian Methods in the Analysis of Linguistic Change. Literary and Linguistic Computing 22(2), pages 137–150.

[20] Fox, J. (2002). Bootstrapping Regression Models Appendix to An R and S-PLUS Companion to Applied Regression.
[21] Canty, A., and Ripley, B. (2021). boot. Bootstrap R (S-Plus) Functions. R package version 1.3-25. Accessible at: https://cran.r-project.org/web/packages/boot/boot.pdf.
[22] Lüdecke, D. (2020). _sjstats: Statistical Functions for Regression Models (Version 0.18.0). Accessible at: https://CRAN.R-project.org/package=sjstats.
[23] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees.
[24] Breiman L. (2001). Random forests. Machine Learning, 45(1), pages 5–32.
[25] Liaw A., and Wiener, M. (2002). Classification and Regression by randomForest. R News 2(3), pages 18–22.
[26] Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A., and Van Der Laan, M. (2006). Survival Ensembles. Biostatistics, 7(3), pages 355–373.
[27] Strobl, C., Boulesteix, A., Zeileis, A., and Hothorn, T. (2007). Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics, 8(25). Accessible at: http://www.biomedcentral.com/1471-2105/8/25.
[28] Strobl, C., Boulesteix, A., Kneib, T., Augustin, T. and Zeileis, A. (2008). Conditional Variable Importance for Random Forests. BMC Bioinformatics, 9(307). Accessible at: http://www.biomedcentral.com/1471-2105/9/307.
[29] Strobl, C., Hothorn, T. and Zeileis, A. (2009). Party on! A New, Conditional Variable Importance Measure for Random Forests Available in the party Package. Department of Statistics: Technical Reports, No. 50.
[30] Schimon, A., and Rabus, A. (2016). Wahrnehmungsdialektologische Untersuchungen zum Russinischen in Zakarpattja am Beispiel der Region Chust. Zeitschrift für Slawistik 61(3), pages 401–432.

sciendo
DOI 10.2478/jazcas-2021-0056

A CORPUS OF CZECH ESSAYS FROM THE TURN OF THE 1900s

PETR POŘÍZKA
Department of Czech Studies, Faculty of Arts, Palacký University, Olomouc, Czech Republic

POŘÍZKA, Petr: A corpus of Czech essays from the turn of the 1900s. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 618 – 630.

Abstract: A literary essay is an interesting unit for language analyses, as its stylistic means often exceed the boundaries of the genre of an artistic essay. The article presents a new corpus of Czech literary essays covering approximately fifty years from 1890 to 1940. Along with the characterisation of the corpus and its annotation, the paper focuses on the TXM corpus tool: In the second part of the study, we use selected texts to conduct an analysis of seven various authors through multidimensional cluster analysis, factorial correspondence analysis and a specificity score. The main parameter of the analyses was usage of parts of speech in texts by individual authors. At present, the Corpus of Czech Essays contains 40 essayist titles written by 15 authors covering various topics (music, visual arts, theatre, literature, etc.).

Keywords: annotation, corpus, corpus linguistics, quantitative analysis, literary essay, multidimensional analysis, orthography, specificity score, TXM

1 INTRODUCTION

At present, Czech linguistics already has a number of corpora available, covering a range of areas with regard to both temporal and typological or genre characteristics. Some textual areas or language periods are, however, covered to a lesser extent or are awaiting processing. One interesting period in the development of standard Czech is the turn of the 1900s, when the views of standard Czech and its orthographic form were established. Attitudes of linguists' changed turbulently during this time. One might mention in this context various polishings representing purist efforts and tendencies, followed by the attempts at stabilization of standard Czech through grammar guidebooks and rulebooks (especially that by J. Gebauer), and finally the Prague Linguistic Circle which regarded the form of standard Czech as one of the key topics.

Czech literary essays from this time illustrate this period of development and also have an indisputable literary-aesthetic value. Although the genre is rather narrowly focused, the options for their utilization for language analyses are undoubtedly wider, since the language and stylistic means used by the authors included in the corpus often exceed the genre of the literary essay. The language of

these texts oscillates, reflecting the means of multiple functional styles: artistic, scientific, journalistic, rhetorical (and partially colloquial); perhaps the only one not involved is the administrative style. This wide range and certain typical tendency to overstep the borders and blend individual functional styles are also confirmed by the only fairly comprehensive anthology of Czech literary essays published in two volumes ([1], [2]). Opelík structures the second volume of the anthology into chapters covering program, portrait, poetological, reprimanding and reflexive essays [2]. The delimitation of the essayist style, as a relatively autonomous unit within the system of functional styles, was first attempted by Havránek in his commentary on functional differentiation of language (1932), although he did not classify it among the basic styles (cf. [3], [4]). Hausenblas [5] classified it as a *complex* style (in contrast to *simplex* styles) and within present-day theory of functional styles, it is classified as a *secondary* style, cf. [6]. When determining the style-based essence of the essay, Jedlička ([4], [7]) pointed out (a) its characteristic tendency to weakening of terminological saturation of a text and (b) a significant proportion of the register of highly formal and dynamic language means. Mistrík [8] defined the borderline character of the literary essay in relation to the (i) scientific, (ii) journalistic and (iii) artistic style. Literary essays are interesting even from the perspective of the lexical means used: formal, expressive and even exclusive means, nonce words, figurative expressions, etc. It was particularly the above-described linguistic character of literary essays – its multifaceted and borderline nature, oscillation among multiple functional styles and mutual blending of language means from various styles – that encouraged us to create a corpus of Czech literary essays (hereinafter also CCE).

2 CHARACTERISTICS OF THE CORPUS

A corpus of this kind must necessarily include texts written by the founder of Czech literary essays F. X. Šalda, the "poet" of Czech essays Otokar Březina, as well as philosophical essays by Ladislav Klíma. The corpus incorporates almost 6 thousand pages of various types of texts (fictional, scientific, journalistic) from various areas (music, visual arts, theatre, literature, etc.). In total, the corpus presently contains 40 books of essays by 15 authors (i.e. on average two to three books for every author) published between 1890–1937. The following authors are included in the present version of CCE: Otokar Březina, Josef Čapek, Karel Čapek, Jaroslav Durych, Otakar Hostinský, Jiří Karásek, Ladislav Klíma, F. V. Krejčí, Jiří Mahen, Miloš Marten, Vilém Mathesius, Arne Novák, Arnošt Procházka, H. G. Schauer and F. X. Šalda.

2.1 Data sources and data processing

The texts included in the corpus come from several sources. The most important one is *Digital Library Kramerius* – a database of the National Library [9]. In

addition, we also used *Digital Library* of the Moravian Library [10], complemented with library loans and OCR conversion of texts into an electronic version. Along with the selection of a particular author and text, the key parameters also included a free license with respect to copyright – expired copyright protection (70+ years from the author's death) – and the version of a particular text: we used the first edition.

For the processing of data, we used the help of students within specialized seminars: each student processed a part of a particular book (ca. 100 pages). Source texts were available in two versions: (1) a set of scanned images (jpg) – the original of the book, and (2) a folder with texts after an automatic OCR conversion (txt). There was a need to make a detailed and precise manual correction for every book according to the original text, as the electronic version (ad 2) is available in *Kramerius* and *Digital Library* databases in a non-revised version – i.e. including all the errors resulting from the automatic conversion. The main editing principle was fidelity to the original. When needed, the text was supplemented with a corrector's note describing a particular change to the text. There was a need, for example, to edit words written in "spaced characters" (a common typographic practice of the particular period), i.e. for the word "u m ě n í" [a r t] (and similar cases elsewhere) it was necessary to delete the spaces between the individual characters and write the expression as "umění" [art]. The whitespace is one of the segmentation characters in corpus databases and without this editing change, the corpus manager would not process these cases as a single lexical unit, but as a sequence of five individual characters "u", "m", "ě", "n", "í". Similarly, there was a need to delete word division of the typographic layout of the book and pagination or add missing signs (typesetting mistakes), for instance "p dstata" [e sence] was corrected to "podstata" [essence] (with an inserted note indicating the missing "o" in the original).

In addition, a list of so-called *anomaly expressions* was purposefully created for every text with regard to the differences between the present-day and contemporary versions of orthography as well as due to the need for linguistic annotation of the texts – for subsequent corrections of automatic annotation (lemmatization and tagging).[1] The usage of this dictionary is wider, however, it allows for insight into the contemporary specific lexicon or the unique lexicon of a particular author (words such as *srostlivost* [a tendency to coalesce], *zvášnivělý* [impassioned], etc.) and may serve as instrument for analyses of texts from the database. The most common 'anomalies' were related to the following phenomena:

- the quantity of vocalic letters: *system* ['system', in Czech correctly "systém"], *primarní* ['primary', in Czech correctly "primární"]

[1] The accuracy or error rate of the annotation depends, among other things, on the tool dictionary. Our comprehensive list of anomaly words that are not part of these annotation dictionaries, may therefore be purposefully used for correction of errors of the automatic text annotation.

- orthographic rules for words of foreign origin, especially Latin and Greek:
  - vocalic digraphs (*aether*) [‘ether’]
  - ending *-ism* (*heroism*) [‘heroism’]
  - double consonant letters *ll* (*illuse*) [‘illusion’], *tt* (*marionetty*) [‘marionettes’], *rr* (*korrelata*) [‘correlates’ – plural noun], *ss* (*associace*) [‘association’], *mm* (*summa*) [‘sum’], *ff* (*affirmující*) [‘affirmating’ – present participle], *kk* (*akkumulace*) [‘accummulation]
  - other phenomena: *th* (*hypothesa*) [‘hypothesis’], *s/z* (*kausalita*) [‘causality’], *ks* instead of *x* (*ekstase*) [‘ecstasis’], *k* instead of *ch* (*karakter*) [‘character’], *qu* instead of *kv* (*quanta*) [‘quantities’].

2.2 Tool for data mining

The main corpus manager for data mining is TXM (abbrev. Textometrie) [11]. This open-source tool was chosen for a number of reasons, for instance the following:

- Unicode – XML & TEI compatible platform
- helps to build various corpus configurations; provides a large spectrum of input formats and rich data models[2]
- has broad and complex options for qualitative-quantitative data mining
- based on the efficient CQP full text search engine and its powerful CQL query language
- has enhanced functions uncommon in other corpus managers[3]:
  - the R statistical environment [12]; provides quantitative analysis, based on R packages (including the option for additional installation of any extension package), e.g.:
    - factorial correspondence analysis
    - hierarchical cluster analysis
    - specific patterns analysis (specificities)
  - includes TIGERSearch query tool for syntactic data mining
  - applies various NLP tools on the fly on texts before analysis (e.g. TreeTagger for lemmatization and POS tagging)
  - provides scripting facilities for repetitive or lengthy tasks automation or for platform extension.

2.3 Corpus format and annotation

CCE was annotated using the open-source tool *MorphoDiTa* [13] which uses a freely accessible Czech morphological dictionary *MorfFlexCZ*[4]. The texts were

[2] For more information, see the documentation of the tool: http://textometrie.ens-lyon.fr/spip.php?rubrique64.

[3] We mean here standard non-commercial corpus managers such as NoSketch Engine, KonText, Poliqarp, etc.

[4] Available at: https://ufal.mff.cuni.cz/morfflex.

lemmatized, morphologically tagged (the Czech 15-position tagset is used – see f.n. 5) and processed into XML format. The corpus annotation is represented in XML through its elements and attributes: directly with *elements* (as a structure *s-attribute*) and/or with *attributes* of these elements (positional *p-attribute*). The basic XML format would therefore look as follows (the root element `text` contains additional metadata – author, title and year of publication of the text)[5]:

```
<?xml version="1.0" encoding="UTF-8"?>
<text author="Durych" title="Essaye" year="1931">
<s>
<w lemma="oheň" pos="N">Oheň</w>
<w lemma="lidstvo" pos="N">lidstva</w>
<w lemma="svítit" pos="V">svítí</w>
<w lemma="dva" pos="C">dvěma</w>
<w lemma="plamen" pos="N">plameny</w>
...
</s>
...
</text>
```

Explanatory note: elements `text` = root element; `s` = sentence; `w` = word; attributes `lemma` and `pos` = part-of-speech.

The second most used corpus format is WPL (word per line) based on column annotation. And if there is a need, the XML format can be converted to vertical WPL-format and imported into the TXM tool using the function *Import CQP* or imported into other corpus managers based on the Manatee system (like (No) SketchEngine and more).

This annotated format was further adjusted: we extracted some of the nominal and verbal sub-categories in order to subsequently use them for corpus analysis. Specifically, a complex tag (`tag`) was used to create separate attributes for the part of speech (`pos`), gender (`g`), number (`n`), case (`c`), person (`p`), and tense (`m`).

Cf. examples below – (1) original annotation from the tool *MorphoDiTa*, and (2) the final form of annotation following adjustments (a sample of Březina's essay *Tajemné v umění* [Mystery in Art]):

(1)

```
<s>
 <w lemma="odpověď" tag="NNFP1-----A----">Odpovědi</w>
 <w lemma="být" tag="VB-P---3P-AA---">jsou</w>
 <w lemma="věčný" tag="AAFP1----1A----">věčné</w>
...
</s>
```

[5] Within the attribute `pos` the individual parts of speech are already referred to with their usual abbreviations of the Czech tagset. For more information, see: https://ufal.mff.cuni.cz/pdt/Morphology_and_Tagging/Doc/hmptagqr.html.

(2)

```
<s>
 <w lemma="odpověď" pos="N" tag="NNFP1-----A----" g="F" n="P"
  c="1" p="-" m="-">Odpovědi</w>
 <w lemma="být" pos="V" tag="VB-P---3P-AA---" g="-" n="P" c="-"
  p="3" m="P">jsou</w>
 <w lemma="věčný" pos="A" tag="AAFP1----1A----" g="F" n="P"
  c="1" p="-" m="-">věčné</w>
...
</s>
```

3 ILLUSTRATIVE ANALYSIS USING THE TXM TOOL

Due to the limited extent of this paper, it is impossible to conduct a more complex analysis, but we shall attempt to illustrate the usefulness of some of the extended options of the TXM tool for corpus analysis of texts. We will move from standard data mining, based on queries for concordances, frequency dictionaries, collocations and other similar phenomena, to multidimensional text analysis.

We randomly selected several texts written by seven authors, particularly:

- Březina – *Hudba pramenů* [Music of the Springs]
- Klíma – *Svět jako vědomí a nic* [World as Consciousness and Nothing]
- Čapek K – *Marsyas* [Marsyas]
- Čapek K – *O umění a kultuře* [On Art and Culture]
- Čapek K – *Kritika slov* [A Critique of Language]
- Čapek J – *Kulhavý poutník* [The Lame Pilgrim]
- Čapek J – *Nejskromnější umění* [The Humblest Art]
- Čapek J – *Co má člověk z umění* [What Man Gets from Art]
- Durych – *Essaye* [Essays]
- Mathesius – *Kulturní aktivismus* [Cultural Activism]
- Šalda – *Boje o zítřek* [Battles for Tomorrow]
- Šalda – *Duše a dílo* [Soul and Work]

The R tool implemented in TXM enables us to use two types of multidimensional analysis: (1) *cluster analysis*, which produces dendrograms expressing similarities or differences between the individual entities compared (text, author, genre, etc.), and (2) *factorial correspondence analysis*.[6] Both these types of quantitative analysis enable comparing the *p-attributes*, i.e. not only lexical items (`word`, `lemma`), but also grammatical categories (part of speech, gender, person, etc.). Apart from lexical analysis, one of the morphological categories showing interesting results using the data sample from CCE is e.g. the grammatical category of number (a tendency for

[6] For more detailed information about both types of analysis, see TXM Manual – ref. [14], pp. 107ff., and reference [15].

usage of singular forms in the case of Šalda, and a very strong tendency for utilization of the plural in the case of Březina). Concerning the parts of speech, an author that differs more significantly from the others is Klíma; the reason for the difference is, however, very specific (see below).

3.1 Ad 1 – Analysis of clusters: Dendrograms

Fig. 1 presents the result of a cluster analysis regarding the main parts of speech for the individual authors.

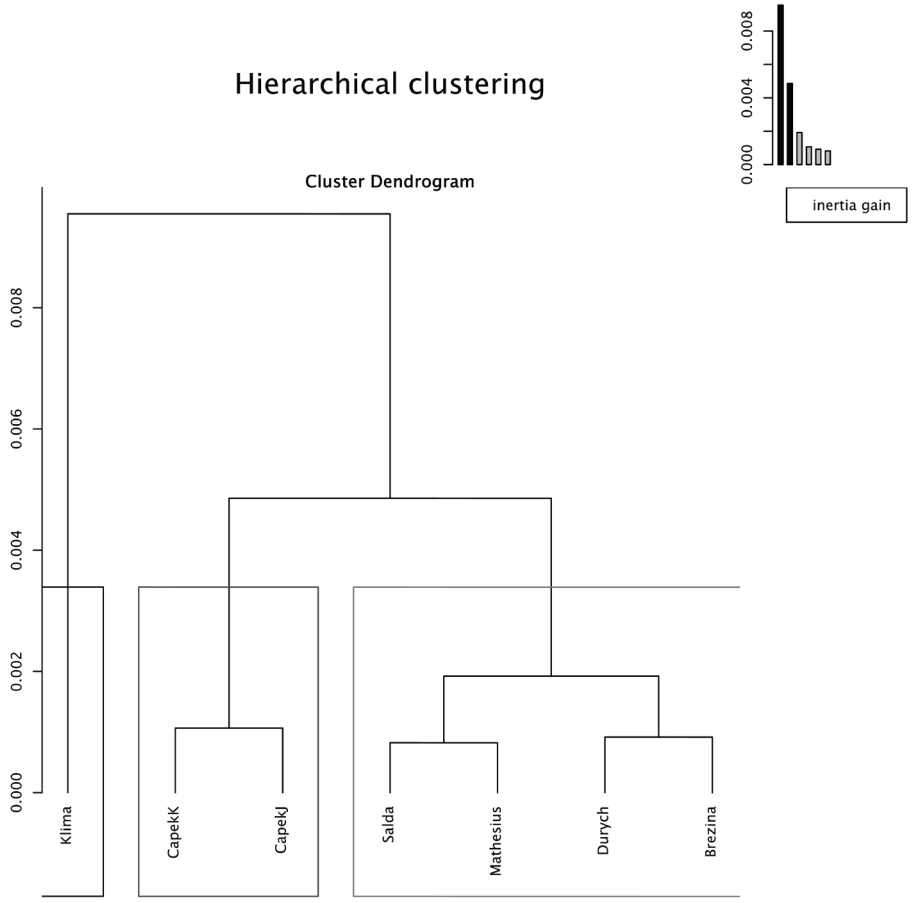


Fig. 1. Dendrogram – the main parameter: part of speech (POS); data: corpus CCE; tool TXM

It is apparent that the biggest difference is that between Klíma and other authors, which is also confirmed by additional three-cluster sub-analyses with the parameters `lemma`, `word`, `pos`, and `tag` (Table 1):

parameter of analysis	cluster 1	cluster 2	cluster 3
`lemma` and `word`	Klíma	Šalda	others
`pos` and `tag`	Klíma	Čapek brothers	others
`n` (grammatical number)	Březina	Mathesius, Šalda, Durych	Klíma, Čapek brothers

Tab. 1. Three-cluster sub-analyses with parameters `lemma`, `word`, `pos`, and `tag`; data: corpus CCE; tool TXM

The reason for the difference in the texts written by Klíma from the rest of the authors is documented in the following correspondence analysis, which also shows the reason for Březina's difference with regard to the grammatical number, which is complemented with a visualization of the specificity score analysis.

3.2 Ad 2 – Factorial correspondence analysis

Using the p-attribute `pos` even for the subsequent correspondence analysis, we can identify a rather specific reason for Klíma's difference: in fact, it is not a POS category, but instead a difference in punctuation (see the tag Z for Klíma):

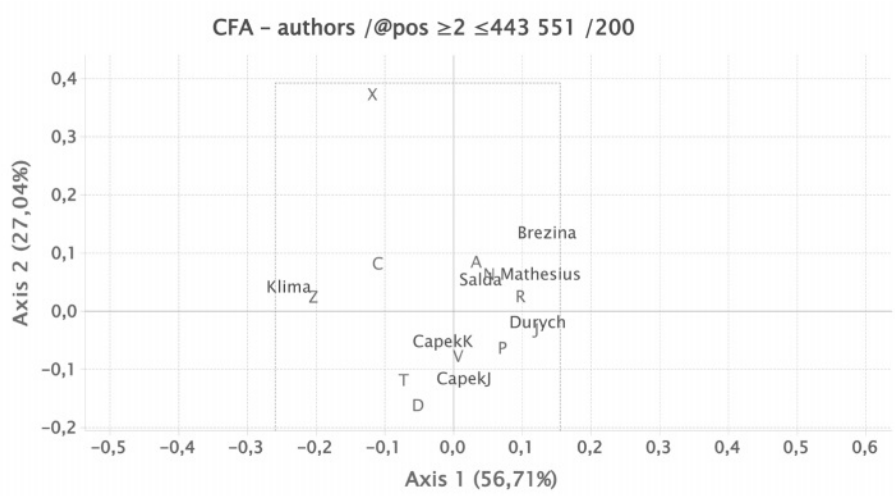


Fig. 2. Factorial analysis – the main parameter: part of speech (`pos`); data: corpus CCE; tool TXM

Klíma's manner of using punctuation marks is highly specific, as illustrated in Fig. 3: m-dashes combined in various ways with a sequence of periods, a semi-colon, or a colon (further combined with a period, a comma, a question mark, or an exclamation mark).


24.

— Nejrozumnější však maxima pro tvory, kteří mohou jednání své vědomě řídit, byla by: „*Dělej cokoli*!“ —: Jest nanejvýš lhostejno, jak jednáš, nejen pro svět, ale i pro tebe! —: jest předem jisto, jak vše dopadne: — — *Každý myšlenkový atom sloučí se průběhem světového roku se* **všemi** *myšlenkovými atomy a* **stane se tak za dobu světového roku tím, čím je svět v každém momentu**; — atom myšlenkový je v témže poměru k světu, jako atom časový k světovému roku. — Dále: *Z výlučné intellektualnosti světa následuje výlučná realnost* **všech** *pomyšlení*, následuje, že není představy bez **realního** důvodu, přání bez **skutečného** objektu; — *že každé pomyšlení jest jen poukazem na skutečnost, každé přání na vyplnění*; — **nic neleží mimo železný kruh vůle v tomto veskrze intellektualním světě**; — nic není irrealního v tomto irrealním světě —. Poněvadž se však každý myšlenkový atom sloučí se všemi ostatními, následuje, **že všechna přání a pomyšlení musí jednou dojít splnění a uskutečnění**! Konečně: Jako positiva a negativa abstracta „svět“ ruší se každým okamžikem, **tak ruší se positiva a negativa každé části každým světovým rokem**: neboť kombinace jednoho atomu se všemi, stejným dílem positivními a negativními, musí být rovněž stejným dílem positivní a negativní; následuje: **svět je vůbec bez hodnoty**... — —: — — Všeho dosáhneš, a vše dosáhne tebe, dělej co dělej! Tvá práce je zbytečna, tvá lenost bez významu, — tak i tak dosáhneš všeho!... *Všechna* tvá nejsmělejší přání se splní, všechny tvé nejhroznější obavy se uskuteční, všechny tvé nejvzdušnější fantasie stanou se realnosti! Povzneseš se k nejvyššímu, klesneš k nejhlubšímu! Nemyslitelné stane se tvou myšlenkou, a nemožné stane se ti skutečností...! Prožiješ všechny metamorfosy, které jsi schopen si představit, a millionkrát více těch, které jsou ti nepředstavitelný!... Ale co budeš mít z tohoto všeho?: — Praničeho! Vším nestaneš se ani šťastnější, ani nešťastnější, ani větší, ani menší... — Všechno tvé úsilné hledání štěstí nerozmnoží ho, všechno tvé odříkání nezmenší ho! Tvé obavy před bolestí jsou nesmyslny: trpíš-li, raduj se, žes si toho zas už kus odbyl; raduješ-li se, věští to pouze bolest, *přesně* tím větší, čím větší tvá radost! Nanejvýš lhostejno jest nejen, co činíš, ale i co se ti přihodí!: jednou musíš si vše odbýt! Tvé úsilí zlepšit tvůj stav je směšné, radost rovněž! každé zlepšení zaplatíš až do posledního haléře zhoršením, štěstí neštěstím, velkost malostí; ale každé tvé špatno promění se v dobro. Nesmíš v *nic doufat*, ale nemusíš se *ničeho obávat*! — Až bude závratný koloběh všeho u konce, jaký bude výsledek...?: *Žádný* ..., — *jako by nic nebylo bývalo*...: *jediná cena všeho* se zatím v koloběhu tom *rozplynula*... A pak začne tento příšerný svět otáčet se znovu..., a po věčnosti věčností... — — Krátce před svou smrtí snil Lichtenberg, že zavítal do vesnické krčmy. Mladý jakýs muž jedl tam polévku, občas vyhazoval ji do výše a lžící opět chytal. Jiní mužové hráli tam v kostky; vedle nich pletla vysoká, hubená žena. Jí ptal se L., možno-li zde co vyhrát? Na odpověď: „nic“ — tázal se pak, možno-li zde co prohrát, a dostal odpověď: „ničeho“. „To považoval jsem za důležitou hru,“ — praví k tomu L. — —: Ano, za důležitou považujeme životní hru, při níž na konec nemůžeme ani ztratit ani získat! Zde je nejvnitřnější *hrozné tajemství* tohoto světa — fantomu: Vše snaží se a plahočí, touží a děsí se, doufá a zoufá, jásá a naříká pro něco, co nejen theoreticky „jest“ *ničím*, ale co i prakticky **v nic** se *paralysuje*...

Fig. 3. Original and specific punctuation of L. Klíma; source: the book *Svět jako vědomí a nic* [The World as Consciousness and Nothing] (1904)

This multidimensional analysis enabled us to detect a highly interesting factor of Klíma's punctuation (worthy of further analysis). It would be appropriate, however, to consider even filtering out this category, which could result in higher precision of the part-of-speech analysis.

A correspondence analysis of the grammatical number also reveals interesting results, where a similar deviation of Otokar Březina from other authors may be observed in the Fig. 4.

A highly useful function that may explain the reason for this obvious difference is the "specificity score" ([14], [16]), which could also be used as one of the alternative approaches to the extraction of prominent text units (thematic expressions,

keywords, etc.).[7] It belongs to the so-called adjusted frequencies which should reflect the actual dispersion or prominence of language expressions or categories in texts, i.e. express their importance rate in the form of hierarchical lists of frequency distribution. TXM even enables a practical option of visualization of these factors, which is helpful when interpreting the results (see Fig. 5).

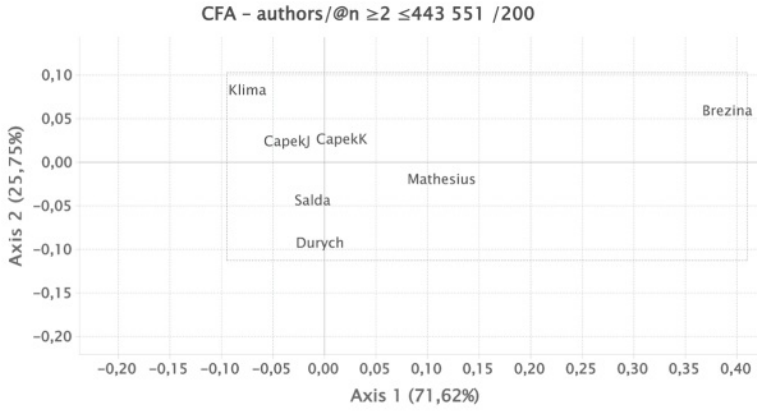


Fig. 4. Factorial analysis – the main parameter: grammatical number (n); data: corpus CCE; tool TXM

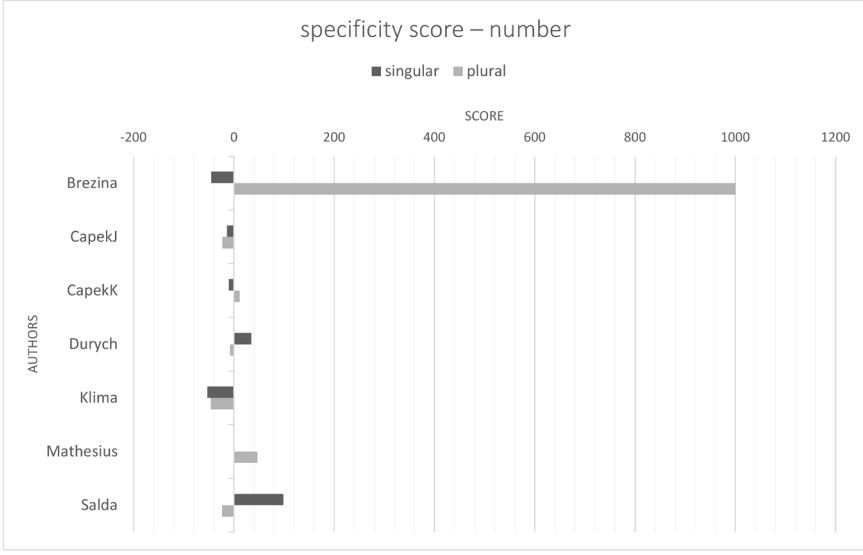


Fig. 5. Specificity score – the main parameter: grammatical number (n); data: corpus CCE; tool TXM

[7] For more information about this quantitative index (including the mathematical formula for its calculation), see TXM Manual [14], pp. 95ff., and reference [16].

The graph clearly indicates a strong, obvious tendency for usage of the plural in the case of Březina (collective plural). We can also observe a very slight tendency for more frequent singular forms in the case of Šalda (subjectivism).

We shall now complement the analysis and the usage of the specificity score with the distribution of the autosemantic parts of speech for the individual authors (we have even added pronouns, as it is an important category for literary texts).

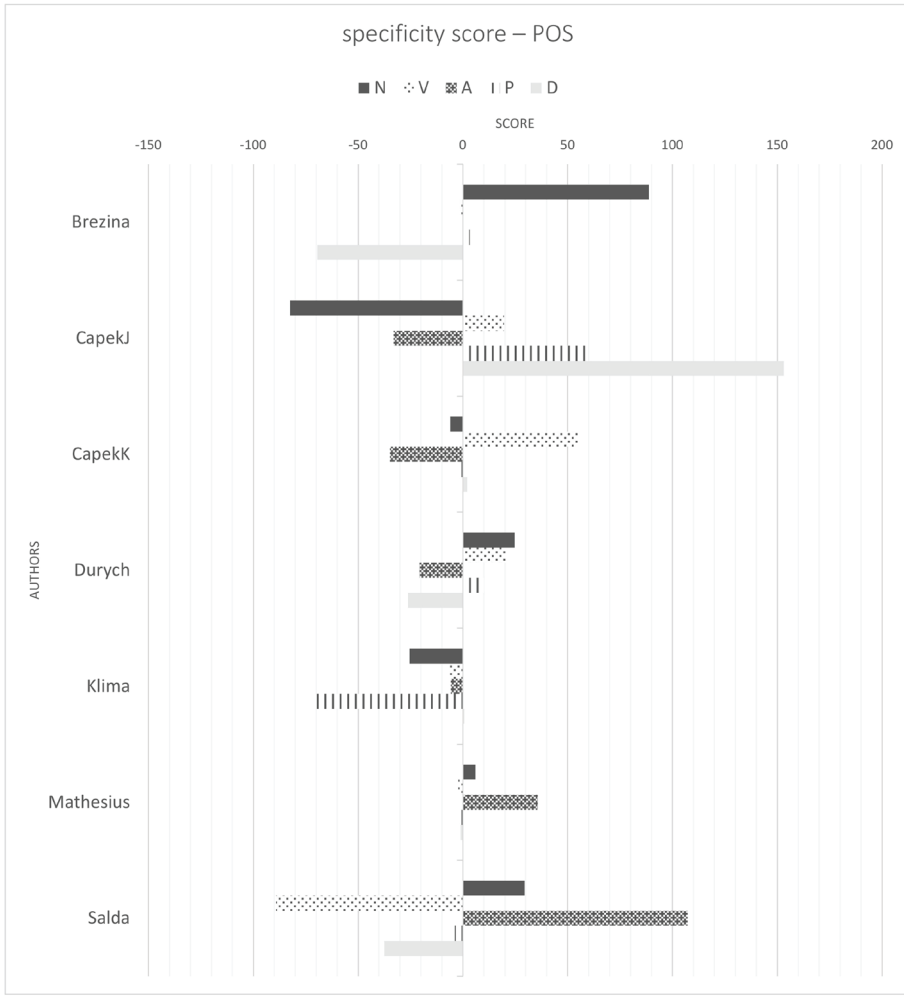


Fig. 6. Specificity score – the main parameter: part of speech (`pos`); data: corpus CCE; tool TXM

Explanatory note: `A` = adjective, `D` = adverb, `N` = noun, `P` = pronoun, `V` = verb.

A visualization of the specificity index reveals the following tendencies in the language of the compared authors:

- The category of nouns is important especially in the case of Březina, to a certain extent also in the case of Durych and Šalda (nominal form of expression, the effort to name substances), in contrast to a rather significant deficit in utilization of nouns in the case of Josef Čapek, compared to the others.
- Adjectives play an important role in the case of Šalda, as well as Mathesius, while a slight deficit is obvious with the Čapek brothers and Durych.
- Pronouns are overused by Josef Čapek, while Klíma suppresses their utilization in his texts.
- Verbs are a dominant part of speech in the case of Karel Čapek, while with Šalda we can see a rather surprising and significant deficit in verbs.
- Adverbs are the most significant part of speech for Josef Čapek, while the opposite tendency may be identified in the case of Březina, and to a certain extent also with Durych and Šalda.

Put simply, we may argue that nouns are the most important and most dominant part of speech in Březina's texts, as with adjectives in the case of Šalda and to a lesser extent also Mathesius. Verbs, a dynamic part of speech, are used to the largest extent by Karel Čapek. In the texts of Josef Čapek, there is a need to focus in greater detail on the prevalence of adverbs, as well as on pronouns. Once again, Klíma is an interesting author: in his texts we can find a deficit in the utilization of nouns and especially pronouns, compared to other authors.

This type of analysis may subsequently serve as background for additional, more traditional, corpus explorations. The findings resulting from this probe enable further analysis to be targeted and focused on more specific phenomena, and especially on those that prove to be relevant or interesting in the texts we are dealing with.

4 CONCLUSION

One of the main aims of the presented project was to establish a linguistically annotated corpus database of Czech literary essays from the turn of the 1900s (we expect that the database will gradually be expanded with new texts and authors). The period from 1890 until the 1930s or 1940s was not chosen randomly: it is a period when the literary essay was formed as a specific, autonomous, and valuable language unit. In addition, the period saw discussions, polemics, and formation of the orthographic form of Czech. This database may therefore serve as a convenient tool for language analyses capturing this development and formation of one language and literary unit.

ACKNOWLEDGEMENTS

The research was supported by the Ministry of Education of the Czech Republic IGA_FF_2020_021 "Czech Studies: Literary and Linguistic Overlaps and Interpretations".

References

[1] Taxová, E. (1985). Experimenty. Český literární esej z přelomu 19. a 20. století. Praha: Melantrich.
[2] Opelík, J. (1986). Lehký harcovník. Antologie českého literárního eseje 2. Léta desátá a dvacátá 20. století. Praha: Melantrich.
[3] Havránek, B. (1932). Úkoly spisovného jazyka a jeho kultura. In Spisovná čeština a jazyková kultura, pages 32–84, Praha.
[4] Jedlička, A. (1989). K jazyku a stylu českých esejistických textů, Slovo a slovesnost, 50, pages 114–125.
[5] Hausenblas, K. (1972). Učební styl v soustavě stylů funkčních, Naše řeč, 55, pages 150–158.
[6] Jelínek, M., and Krčmová, M. (2017). Esejistický styl. In CzechEncy – Nový encyklopedický slovník češtiny. Available at: https://www.czechency.org/slovnik/ESEJISTICKÝ STYL.
[7] Jedlička, A. (1973). K vymezení a charakteristice esejistického stylu, Studia Slavica Pragensia, pages 167–178.
[8] Mistrík, J. (1974). Esejistický štýl, Slovenská reč, 40, pages 321–332.
[9] Digital library Kramerius: Available at: http://kramerius.nkp.cz.
[10] Digital Library MZK: Available at: http://www.digitalniknihovna.cz/mzk.
[11] Textometrie project: TXM (version 0.8.1) [Software]. Available at: http://textometrie.ens-lyon.fr.
[12] The R Project for Statistical Computing: R (version 4.0.4) [Software]. Available at: https://www.r-project.org/.
[13] Morphological Dictionary and Tagger: MorphoDiTa [Software]. Available at: http://lindat.mff.cuni.cz/services/morphodita/.
[14] TXM Reference manual (v0.7). Available at: http://textometrie.ens-lyon.fr/files/documentation/TXMManual 0.7.pdf.
[15] Benzécri, J.-P. et al. (1973). L'analyse des correspondances. Paris: Dunod.
[16] Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus, Mots, 1, pages 127–165.

sciendo
DOI 10.2478/jazcas-2021-0057

BUILDING CZECH TEXTBOOK CORPORA (UcebKo) FOR WORD-FORMATION RESEARCH OF CZECH AS A SECOND LANGUAGE

ADRIANA VÁLKOVÁ
Masaryk University, Brno, Czech Republic

VÁLKOVÁ, Adriana: Building Czech textbook corpora (UcebKo) for word-formation research of Czech as a second language. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 631 – 640.

Abstract: This work-in-progress paper presents a specialized language corpus UcebKo built from textbooks of Czech for foreigners. The corpus integrates three subcorpora (UcebKo-A2, UcebKo-B1, and UcebKo-B2) which allow research of Czech as a second/foreign language at chosen language levels (A2, B1, and B2). In this case, the research is focused on word-formation, where the first results, i.e., mapping of derived words denoting persons, illustrate the approach and methodology used.

Keywords: word-formation, derivational morphology, textbook corpus, Czech as a second language, names of persons

1 INTRODUCTION

Most language corpora can be understood as extensive databases of parts of texts, in which it is possible to search and sort individual text units (sentences, word combinations, words, etc.) and observe them in their natural context ([1]). For most linguistic research, it is more appropriate to work with the so-called annotated corpus, where each corpus text unit (the so-called token) is provided with a lemma, the word form itself, and morphological information in the form of tags (i.e., information about the part of speech and its grammatical categories). In terms of general vs. specialized (specific) lexicon we differentiate between two types of corpora – general and specialized. General corpora are built for the sake of making generalizations (relating to morphology, lexicology, etc.) about the language. Specialized language corpora (in contrast to general corpora) always have a specifically defined purpose for which they are built – there are many types of specialized corpora (e.g. [2]).

In the case of specialized corpora made up of textbooks, the so-called textbook corpora (e.g., [3], [4]), the aim can be twofold – 1. to map the language (metalanguage) of textbooks, i.e., the linguistic and/or pedagogical research is aimed at all parts of the textbook or 2. to capture the vocabulary of the target group of textbooks. We have aimed at point two, i.e., to build a textbook corpus that would represent the Czech language of foreigners. A corpus like this can be understood as a simplified natural language (specifically the Czech language), where the degree of

simplification is determined by the language level (from A2 to B2). Corpus UcebKo could work as a basis of any research of Czech as a second language. In our case, the corpus works as a basis of word-formation research in which we want to map and then obtain (the most common) affixes for language levels A2, B1, and B2 and present them by using appropriate vocabulary.

2 MOTIVATION

Morphology plays a key role in the acquisition of Czech as a second language, as in other languages with a richly developed morphology (more than 75% of the Czech lexicon consist of derived words, see [5]). In addition to the inflectional morphology, which is focused on creating different forms of one word (e.g., from the word *otec* 'father' forms like *otcovi* 'to the father', *otcové* 'fathers') and which the student – a foreigner – encounters from the beginning, derivational morphology exists as a separate part of word-formation. Derivational morphology deals with the formation (or reproduction) of new words from already existing words (e.g. *mluvit* 'to speak' → *mluvčí* 'speaker'). As both morphologies are closely related – thanks to the suffix from which the word has been derived, it is possible 1. to identify the part of speech and 2. to classify the word within its paradigm (e.g., *cestovatel* 'traveller' is derived by the means of the suffix *-tel*, i.e., it is a noun which is inflected according to the *muž* 'man' paradigm). Some of the word-formation rules in Czech are mostly well acquired (e.g., adverbialization of adjectives: *krásný* 'beautiful' → *krásně* 'beautifully'), but most of them cause problems due to 1. the polyfunctionality of most suffixes (e.g., the suffix *-ka* with about 27 different meanings: a person (*manželka* 'wife'), an appliance (*sušička* 'a dryer'), a deminutive (*dcerka* 'little daughter') etc., 2. the irregular morphological alternations (e.g., *e/a*: *vejce* 'an egg' → *vaječný* 'made from eggs', [6]) and due to 3. many options of how to name the facts around (e.g., in Czech there are about 19 suffixes for naming a person according to the action which this person does).

There is no publication or textbook that systematically works with word-formation, or the existing textbooks do not provide a complete view of the word-formation system of Czech although their vocabulary could be many times more extensive if the students-foreigners acquired the word-formation principles in Czech.

We assume the results of this corpus research could be useful as a basis for any work with word-formation or for any word-formation project intended for students-foreigners.

3 TEXTBOOK CORPUS UcebKo

3.1 Corpus characteristics and composition

The UcebKo corpus is a specialized language corpus created from nine textbooks of Czech for foreigners including the keys thereto (that is, where the key

was available). This type of corpus represents the vocabulary that should be acquired by students-foreigners (contrary to the learner corpus which represents vocabulary that has been already acquired by students-foreigners, including the acquired mistakes). In general, UcebKo represents a sample of natural language (Czech) which is simplified based on the certain language level (A2–B2). It is possible to assume that textbooks capture rather the core of the lexicon than its periphery (which must be taken into account for linguistic research).

UcebKo integrates three subcorpora:

1. **UcebKo-A2** created from Czech textbooks for foreigners for level A2,
2. **UcebKo-B1** created from Czech textbooks for foreigners for level B1,
3. **UcebKo-B2** created from Czech textbooks for foreigners for level B2.

The designations according to the CEFR (see [7]) were used in all three corpora mentioned, in accordance with the textbooks they have been derived from. Level A1 was intentionally omitted, because according to a search of the textbooks of Czech for foreigners, they do not deal with word-forming phenomena at such a low language level (an exception is just a single textbook).

Every subcorpus always consists of three textbooks (see Table 1). In general, those authors who have written a textbook for more than one language level were preferred.

	UcebKo-A2	**UcebKo-B1**	**UcebKo-B2**
1	*Česky krok za krokem 1* ‘Czech Step by Step 1’ (from 13th chapter)	*Česky krok za krokem 2* ‘Czech Step by Step 2’	*Čeština pro azylanty a cizince (B2)* ‘Czech for asylum seekers and foreigners (B2)’
	L. Holá (2016) Praha, Akropolis	L. Holá – Bořilová, P. (2014) Praha, Akropolis	A. Adamovičová et al. (2006) Brno, SOZE
2	*Česky, prosím II.* ‘Czech, please II.’	*Česky, prosím III.* ‘Czech, please III.’	CZech it UP! B2
	J. Cvejnová (2012) Praha, Karolinum	J. Cvejnová (2016) Praha, Karolinum	D. Hradilová (2020) Olomouc, UPOL
3	*Čeština pro cizince A1 a A2* ‘Czech for foreigners A1 and A2’ (from 5th chapter)	*Čeština pro cizince B1* ‘Czech for foreigners B1’	*Čeština pro cizince B2* ‘Czech for foreigners B2’
	M. B. Kestřánková et al. (2017) Brno, Edika	M. B. Kestřánková et al. (2016) Brno, Edika	M. B. Kestřánková et al. (2020) Brno, Edika

Tab. 1. Composition of corpus

3.2 Corpus size

The UcebKo corpus spans 303,862 words and the size of each subcorpus is different (see Table 2). The smallest is the UcebKo-B2 subcorpus and the largest is the UcebKo-B1 subcorpus.

	UčebKo-A2	UčebKo-B1	UčebKo-B2
number of words	91,561	122,604	89,697
number of sentences	15,099	16,993	9,739
average sentence	6 words / sentence	7 words / sentence	9 words / sentence

Tab. 2. Size of the UcebKo corpus

The size of an average sentence (in terms of the number of words the sentence consists of) differs for each level (see 3rd line in Table 2) – data have been found thanks to statistical data of the corpus interface. An average sentence at the B2 level is 9 words, which is quite a long sentence and, therefore, it can be assumed that complex sentences will predominate.

3.3 Criteria for textbook selection

The textbooks from which the corpora are made up have been chosen according to the six criteria that were set with regard to the purpose for which the corpora have been built:

1. **criterion determines the target group of textbook users – adults** (because the results of corpus research will be used in projects which are primarily intended for adult students-foreigners),
2. **criterion is the type of textbook should not be purely grammatical but should be more conversational** (because we want to capture as much of the natural context of the derived words as possible, not the grammatical rules, etc.),
3. **criterion is the recency of the textbook and it was decided not to incorporate a textbook older than 15 years** (because we want to work with the most recent Czech lexicon),
4. **criterion is the number of textbooks – three textbooks for each subcorpus**. This criterion was chosen due to impossibility to obtain more than three textbooks for the B2 language level. Distinction between B1 and B2 allows word-formation research according to language levels. Moreover, we presume the most frequent words (the core of the lexicon) do not change with a larger corpus size (cf. same as in general Czech corpora),
5. **criterion is that each subcorpus consisted of textbooks from different authors** (because we assume, a corpus built from textbooks of more authors is more objective than a corpus built from textbooks of only one author),
6. **criterion is that the textbooks were actually used in practice**, i.e., in teaching, etc.

3.4 Access to corpus

The corpus is accessible at the website of Sketch Engine (https://ske.fi.muni.cz/) only for verified users who have submitted a statement in which they undertake to use the corpus just for research purposes and also to publish corpus parts with only complete citations.

4 BUILDING OF THE UcebKo CORPUS

4.1 The corpus-building process

The textbook corpus UcebKo was created in SketchEngine ([8]), a corpus interface, thanks to which an already annotated corpus was created (Sect. 1). All Czech corpora created in SketchEngine have been annotated by the morphological analyzer Majka ([9]) and Desamb ([10]). The process of corpus building can be described as the result of five steps:

- **Obtaining the textbooks**.
- **Textbook scanning**.
- **OCR / Copy text**: The textbooks were converted from their scanned form to plain text using a program with the OCR function.
- **Cleaning text**: In this step, it was necessary to 1. check the obtained text against the original text from the textbooks and 2. set the criteria for what to keep in the text and what to remove (more info in 4.2 Cleaning text).
- **Creating corpus**: The text document was uploaded to the corpus interface which created the corpus automatically. This process takes between several seconds and a maximum of several minutes, so this is why it is the easiest step in the whole process of building a corpus.

4.2 Cleaning the text

First, it was necessary to check the obtained text against the original text in the textbook, because different errors may have occurred during the scanning and/or OCR phase. There were errors especially like bad text recognition due to a graphically processed background on which the text was written or incorrectly recognized diacritics over some words etc. After checking the texts, criteria were set by which it was determined exactly what would be kept in the text and what would be removed from the text. It was necessary to clarify the aim for which the corpus was being built in order to determine the criteria for cleaning the text, i.e., to create a database of texts which would represent simplified Czech corresponding to a certain language level. Therefore, only whole sentences were kept in the text and the word combinations or free-standing words were removed. Next, the language of mediation, inscribed pronounciation, and grammar explanation was removed because it does not represent natural language (natural context of words) and finally, the examples of poems were removed because they do not usually reflect current lexicon.

5 WORD-FORMATION RESEARCH IN UcebKo

5.1 Aim of word-formation research

The aim of word-formation research is 1. to map the quantity of the words derived from suffixes that denote the persons in the vocabulary of foreigners and 2. to capture the suffixes (the so-called word-forming types) from which these words are derived at

language levels A2, B1, and B2. It can be assumed that productive word-formation types will occur across all language levels. However, the aim is to find out which word-formation types are involved. The resulting data will provide a view on the Czech lexicon from the word-formation perspective presented by concrete numbers.

5.2 Methodology of word-formation research

The described research could be understood as a process that consists of three steps: 1. obtaining suffixes from books focused on Czech word-formation, 2. searching for words derived from these suffixes in the corpus, 3. final (quantitative) analysis of the found data.

The starting point of word-formation research was to obtain suffixes denoting a person. The list of suffixes was obtained by searching dictionaries or grammar specializing in Czech word-formation ([11] and [12]). Words containing the suffixes were searched for in the corpus by two queries – first, by a query containing morphological tags and then a query without morphological tags, where the string of characters was searched for at the end (e.g., *-tel*). The second query was performed as a check that takes into account the expected error rate of automatic natural language processing. The error rate found is mainly based on the basic properties of natural language, which is 1. the linguistic homonymy (e.g., the word *mluvčí* 'speaker' was not found by tags specifying the noun), and the fact that 2. the word form is missing in the dictionary of morphological analysis (see [13], [14]) (e.g., the word *antitalent* 'dullard' was not found by tags specifying the masculine animate noun).

After this corpus searching, all masculine animate nouns were searched and analyzed by their endings. In this way, the suffixes *-ál* (e.g. *profesionál* 'professional'), *-át* (e.g., *adresát* 'addressee'), *-eň* (e.g., *vězeň* 'prisoner') were found. The found suffixes were subsequently searched for in the largest dictionary of affixes for Czech (see [15]), in which they were found.

5.3 Derived words denoting persons

A total of 13,486-word forms (1,564-lemmas) were found by using the morphological tags for the masculine animate nouns. However, this number includes 1. a group of words which were incorrectly assigned as animate masculine (e.g., *knedlík* 'dumpling') and 2. words that do not denote persons (e.g., *pták* 'bird') which are necessary to remove for analysis. Also, proper individual names (first names, e.g., *Adam* and surnames, e.g., *Novotný*) have been removed, because they do not, in contrary to the rest of the words, denote persons according to certain circumstances or characters. Thanks to manual analysis it was found:

- 643-lemmas (8,350-word forms) were found as words denotating persons,
- out of 643-lemmas, 78.5% (505-lemmas) of lemmas were derived from other word(s),
- out of 505-lemmas, 91.9% (464-lemmas) of lemmas were derived from suffixes.

It was necessary to separate the derived words from loanwords adapted into Czech by suffixes. These loanwords (containing suffixes) constitute 14.8% (95-lemmas) of a total of 643-lemmas. These words are also mentioned here because they could be easily acquired by foreigners with a knowledge of word-formation principles (suffixes).

The number of words denoting persons and the words derived from suffixes differ in the individual subcorpora of UcebKo:

- **UcebKo-A2**: contains 263-lemmas denoting persons, of which 78.5% are derived,
- **UcebKo-B1**: contains 401-lemmas denoting persons, of which 77.1% are derived,
- **UcebKo-B2**: contains 409-lemmas denoting persons, of which 80.1% are derived.

5.4 Identified suffixes with the meaning denoting a person

A total of 28 suffixes with the meaning denoting a person have been found in the UcebKo corpus (see Table 3). Suffixes are sorted by their relative frequencies ([16]) because the work with absolute frequencies is not possible due to the different size of each subcorpus. Most of these suffixes are represented in all three textbook subcorpora, i.e., UcebKo-A2, UcebKo-B1 and UcebKo-B2 (see numbers 1–18 in Table 3). However, the number of suffixes found in the subcorpora is different:

- in A2 21 suffixes were found,
- in B1 20 suffixes were found,
- and in B2 25 suffixes were found.

Some suffixes (marked by *) have been found with a word-formation function (e.g., *student* ‘student’ ← *studovat* ‘to study’) and/or with a lexical function (e.g., *pacient* ‘ patient’ > lat. ‘patiēns‘), it depends on the concrete words. Suffixes with lexical function are part of loanwords and in this sense, the suffixes work as formal instruments of adaptation into Czech. Loanwords are not (naturally) included in the word-formation analysis presented in Table 3.

		UcebKo	**UcebKo-A2**	**UcebKo-B1**	**UcebKo-B2**
	suffix	number of lemmas with this suffix (and their relative frequency)			
1	*-tel*	**68 (7251.2)**	14 (2337.2)	22 (2316.4)	32 (2597.6)
2	*-ik/-ník*	**117 (4516.3)**	28 (1157.6)	40 (1541.5)	49 (1817.2)
3	*-ec (-ovec, -inec)*	**79 (4436)**	18 (1376.1)	31 (1566.0)	30 (1493.9)
4	*-č*	**27 (3847.8)**	7 (1190.4)	11 (1386.5)	9 (1270.9)
5	*-ař*	**29 (2305.8)**	7 (709.9)	14 (1150.0)	8 (445.9)
6	*-ent**	**14 (2297.9)**	3 (819.1)	6 (709.6)	5 (769.2)

		UcebKo	UcebKo-A2	UcebKo-B1	UcebKo-B2
	suffix	number of lemmas with this suffix (and their relative frequency)			
7	*-ista*	**49 (1936.3)**	15 (600.6)	20 (856.4)	14 (479.3)
8	*-ce*	**44 (1439.3)**	10 (382.2)	18 (522.0)	16 (535.1)
9	*-ář*	**56 (1234.9)**	10 (207.5)	20 (481.2)	26 (546.2)
10	*-ák*	**33 (1010.6)**	6 (207.5)	15 (424.1)	12 (379.0)
11	*-an*	**35 (794.7)**	9 (207.5)	14 (252.8)	12 (334.4)
12	*-ik**	**27 (702.7)**	8 (196.5)	11 (261.0)	8 (245.2)
13	*-ér**	**13 (616.3)**	3 (283.9)	5 (187.5)	5 (144.9)
14	*-or (-tor, -átor)**	**20 (330.2)**	4 (54.6)	9 (228.3)	7 (312.1)
15	*-ant*	**17 (288.9)**	5 (65.5)	5 (89.7)	7 (133.7)
16	*-a**	**10 (250.7)**	2 (43.6)	5 (73.4)	3 (133.7)
17	*-íř**	**3 (219.2)**	1 (54.6)	1 (97.8)	1 (66.8)
18	*-ina (-otina)*	**4 (96.9)**	1 (10.9)	1 (57.0)	2 (29.0)
19	*-áč*		-	-	3 (200.6)
20	*-eň*		-	-	1 (89.1)
21	*-át**		-	2 (16.3)	1 (44.5)
22	*-ál*		-	1 (8.1)	1 (22.2)
23	*-oun*		-	-	1 (22.2)
24	*-ita**		1 (21.8)	-	-
25	*-án**		-	-	1 (11.1)
26	*-och*		-	-	1 (11.1)
27	*-ka*		1 (10.9)	-	-
28	*-l*		1 (10.9)	-	-

Tab. 3. Suffixes for derivation of words denoting a person found in UcebKo

As it is possible to see (Table 3), the words denoting persons are most often derived from the suffixes *-tel* (e.g., *učitel* 'teacher'), *-ík/-ník* (e.g., *mladík* 'a young man', *zákazník* 'customer'), *-ec* (e.g., *cizinec* 'foreigner') and *-č* (e.g., *rodič* 'parent') at language levels A2–B2, and moreover, the suffix *-tel* was found to be the most frequent. It must be said that these suffixes, which occur across all subcorpora, play a key role in Czech language acquisition because students-foreigners are confronted with them almost all the time when learning Czech (from A2 to B2).

6 CONCLUSION AND FUTURE WORK

The submitted paper has presented a process of building a textbook corpus called UcebKo, which has been built with the purpose of having language material for word-formation research in Czech as a second language. The corpus UcebKo integrates three subcorpora (UcebKo-A2, UcebKo-B1, and UcebKo-B2) which

allow conducting a separate research of the Czech vocabulary of foreigners at language levels A2, B1, and B2. The described research was focused on the mapping of suffixes from which words denoting names for persons are derived.

The process of building the corpus was described as a five-step process: 1. to obtain textbooks, 2. to scan them, 3. to do an OCR to get the text alone, 4. to clean the text and 5. to create the corpus in the corpus interface of SketchEngine. The cleaning of the text was found to be the most arduous phase of the corpus building. In this phase, it was necessary to decide what to remove (structures without sentence form, the mediation language, the written pronunciation, the examples of poems, and grammar explanations) and what to keep (only the sentence structures that are not grammatical in nature only).

The suffixes from which the words denoting persons are derived at language levels A2, B1, and B2 have been found and presented in the form of lists. Eighteen suffixes have been found at all researched language levels: *-a, -ák, -an, -ant, -ař, -ář, -ce, -č, -čí, -ec, -ent, -ér, -ián, -ik, -ík, -ina, -íř, -ista, -or (-tor/-átor), -tel*. Moreover, it was found that most often persons are named by the suffixes *-tel, -ík/-ník, -ec*, and *-č*. The data could be compared with data from general corpora in the future (but it will be undertaken as individual research due to the large polyfunctionality of the suffixes).

The next research will be focused on describing the meanings of the found derivatives by their semantic features. The result will be presented in lists intended for language levels A2, B1, and B2. Derived words processed in this way could be useful in the field of didactics of Czech as a foreign language – for lecturers of Czech (for creating word-formation exercises) and for students-foreigners (such as knowledge of how to name a person in concrete circumstances).

ACKNOWLEDGEMENTS

This work was supported by grant No. TL003000293 (Word-formation Analysis Software Tool for Teaching Czech for Foreigners) through the Technology Agency of the Czech Republic.

References

[1] Oliva, K., and Doležalová, D. (2004). O korpusu jako o zdroji jazykových dat. In Korpus jako zdroj dat o češtině. Brno, Masarykova univerzita, pages 7–10.

[2] Cvrček, V. (2021). Struktura Českého národního korpusu. In Wiki Český národní korpus. Accessible at: https://wiki.korpus.cz/doku.php/cnk:struktura.

[3] Vališová, P. (2013). Učebnicový korpus a jeho využití pro výuku češtiny jako cizího jazyka. In J. Klímová, Gramatika a korpus 2012: 4. mezinárodní konference. Hradec Králové. Accessible at: http://utkl.ff.cuni.cz/~rosen/public/GC2012/Konferencni_prispevky/ValisovaPavlina.pdf.

[4] Meunier, F., and Gouverneur, C. (2009). New types of corpora for new educational challenges: collecting, annotating and exploiting a corpus of textbook material.

[5] Dokulil, M. (1962). Tvoření slov v češtině. 1, Teorie odvozování slov. Praha, Nakladatelství Československé akademie věd.
[6] Ševčíková, M. (2018). Modelling Morphographemic Alternations in Derivation of Czech. The Prague Bulletin of Mathematical Linguistics, 110, pages 7–42. Accessible at: https://ufal.mff.cuni.cz/pbml/110/art-sevcikova.pdf.
[7] Ivanová, J. (2002). Společný evropský referenční rámec pro jazyky: jak se učíme jazykům, jak je vyučujeme a jak v jazycích hodnotíme. Olomouc, Univerzita Palackého v Olomouci.
[8] Kilgarriff, A., Rychlý, P., Jakubíček, M., Rundell, M. et al.: Sketch Engine [Computer Software and Informatiom Resource]. Accessible at: http://www.sketchengine.co.uk.
[9] Jakubíček, M., Kovář V., and Šmerk, P. (2011): Czech Morphological Tagset Revisited. In A. Horák, P. Rychlý (eds.), Proceedings of Recent Advances in Slavonic Natural Languages Processing. Brno: Tribun EU, 2011, pages 29–42, 14 p. ISBN 978-80-263-0077-9.
[10] Šmerk, P. (2008): K morfologické desambiguaci češtiny. Accessible at: https://is.muni.cz/auth/th/wteg5/teze.pdf. Advanced Master's thesis. Masaryk University, Faculty of Informatics.
[11] Štícha, F. et al. (2013). Velká akademická gramatika spisovné češtiny. Praha, Academia.
[12] Karlík, P., Nekula, M., and Pleskalová, J. (2016). Nový encyklopedický slovník češtiny. Praha, Nakladatelství Lidové noviny. Accessible at: https://www.czechency.org/slovnik/.
[13] Osolsobě, K. (1996). Algoritmický popis české morfologie a strojový slovník češtiny. Brno, Masarykova univerzita. Disertační práce.
[14] Brno Morphological Analyzer Ajka. Accessible at: https://nlp.fi.muni.cz/projekty/wwwajka/.
[15] Šimandl, J. (2016). Slovník sufixů užívaných v češtině. Praha, Univerzita Karlova, Karolinum. Accessible at: http://www.slovnikafixu.cz.
[16] Kováříková, D. (2021). Frekvence. In Wiki Český národní korpus. Accessible at: https://wiki.korpus.cz/doku.php/pojmy:frekvence.

INTERDISCIPLINARY RESEARCH BASED ON CORPORA

sciendo
DOI 10.2478/jazcas-2021-0058

ON CONCEPTUAL AND AXIOLOGICAL ASPECTS OF THE WORD *MUTTER* 'MOTHER' IN CONTEXT (BASED ON CORPUS MATERIAL)

ANITA BRAXATORISOVÁ
Department of German Studies, Faculty of Arts, University of Ss. Cyril and Methodius, Trnava, Slovakia

BRAXATORISOVÁ, Anita: On conceptual and axiological aspects of the word *Mutter* 'mother' in context (Based on corpus material). Journal of Linguistics, 2021, Vol. 72, No 2, pp. 643 – 655.

Abstract: This paper focuses on a linguistic image of mother in German languages. It seeks to grasp it through a typical context of the German word *Mutter* 'mother'. The research is based on results of distributional and thematic analyses of these words. These analyses are used as a base for reconstructing prototypical characteristics of "mother" and the related concepts used by speakers of German. The paper develops these findings into compiling the most frequent collocations and other (mostly contextual) information gathered by the use of corpus tools. The paper concludes with an outline of unconscious axiological processes used in evaluating the image of mother on the good/bad axis.

Keywords: corpus linguistics, collocation profile, context, linguistic image of a word, axiology

1 THE CONTENT OF THE WORD *MOTHER*

A wide range of connotations and associations of the studied word proves that it is a complex, emotionally varied notion. Its complexness is determined by the fact that this concept has biological (genetic, physiological), psychological, emotional, moral- ethical and social parameters that have been gradually mirrored – as modern linguistics progressed – in ever expanding definitions and descriptions of words. As part of the structuralist explanation it was the biological, or genealogical, basis of the word defined as a direct female ancestor from the previous generation [1].

For Lakoff [2], the meaning of this word is more complex than what can be expressed on a basis of necessary and sufficient conditions applied by the Structuralists. In his works, this author ([2], [3]) describes its meaning as a category whose central model is characterised by certain sub-models:

(1) the model of birth: the mother gives birth to a child;
(2) the genetic model: the mother is woman from whom her child gets one half of its genes;
(3) the model of upbringing and care: the mother is a person who brings up a child and takes care of it;

(4) the model of marriage: the mother is the wife of the child's father;
(5) the genealogical model: the mother is the closest female ancestor of a child.

The author ([2], [3]) is aware of the fact that not all above-mentioned models are in all cases represented by the same person. The expressions such as genetic mother, surrogate mother, stepmother, adoptive mother etc. are used for such cases. He emphasises that the cases when all models can be applied are ideal and prototypical whereas he calls the variants such as genetic mother, surrogate mother, stepmother, and adoptive mother segments of the prototype.

According to Kiefer [4], the above-mentioned expressions prove that the meaning of the word mother covers exclusively the case of a child-bearer; this can change only with the context, including the above-mentioned combinations with attributes such as step-, surrogate etc. According to this author, the meaning of the word mother can be determined on the basis of necessary and sufficient conditions but we must distinguish between the semantics of the word and its use; however, the prototype theory does not take this into account.

Anna Wierzbicka [5], the founder of the natural semantic metalanguage, likewise acknowledges that it is a complex notion but she has nevertheless harshly criticised Lakoff's claim that the notion of mother is psychologically and cognitively so complex that necessary and sufficient conditions of structural semantics are not satisfactory for the description of its meaning. In the definition of the word mother (X is the mother of Y), Wierzbicka distinguishes between:

– the biological level of the meaning where we must, according to her, take into account the real state of affairs [5]:
  "(a) at one time, before now, X was very small
  (b) at that time, Y was inside X
  (c) at that time, Y was like a part of X" [5, p. 155]:

– the psychological-social level of the meaning that is linked to certain expectations by the society [5]:
  "(b) because of, people can something like this about X: 'X wants to do good things for Y
  X doesn't want bad things to happen to Y'" [5, p. 155].

Orgoňová and Bohunická [6] point out two divergent linguistic approaches to defining meanings of words. On one hand, it is the structural-semantic approach that operates with necessary and sufficient conditions of the meaning. Following this approach, these authors define the meaning of the word mother as *"a woman who has (at least one) child"*, or *"a woman who gave birth to a child / passed on genetic material / has a child"* [6, p. 35]. On the other hand is the ethno-linguistic approach

which, based on entrenched stereotypes, associations, connotations, and figures of speech such as collocations or idioms, sees the most important component of the meaning of the word mother to be *"a woman who brings up a child"* [6, p. 35].

Another Slovak author dealing with conceptualisation of the notion of mother was Kmecová [7], who studied it as part of her research focusing on the linguistic image of a woman in the Slovak and Slovenian languages. This author gathered and thematically compiled idioms and non-idiomatic figures of speech based on various physiological, psychological, and social changes in various phases of a woman's life, out of which pregnancy, childbirth, and a relationship with her child are linked to the concept of mother. Furthermore, she has studied equivalence and motivation for creating idioms. She also uses linguistic, historic, social sciences, and ethnographic impulses for uncovering converging and varied ways of perceiving the idiomatic image of the woman and mother in the Slovak and Slovenian languages. The author categorised the collected phrasemes and non-idiomatic expressions related to motherhood into the following topical groups:

"(a) biological-physiological preconditions for motherhood;
(b) the physical and mental condition of a woman during pregnancy;
(c) the social status of a child-bearer, labour and its process;
(d) allusions to the process of lactation and breast-feeding;
(e) the relationship between the mother (stepmother) and a child;
(f) the social status of an out-of-wedlock child" [7, p. 62].

The results of a corpus and discourse analyzes from Černá and Čech [8] focused on motherhood provide a significant stimulus for the content analysis of the concept. Research of the lemma "motherhood" through corpus tools and methods like T-scores and CARDS has shown that *"the chief discourse-of-motherhood categories were: surrogate motherhood, relationship between motherhood and career, delight from motherhood, family relationships, financial and time aspects of motherhood, changes due to motherhood, and active motherhood"* [8, p. 252]. The biological, social, economic and legal problems of the career-motherhood relationship have proved to be topical.

Bańczerowski, a Polish hungarologist, uses specific approach ([9], [10], [11]). He describes the notion of mother – using the terminology of Langacker's cognitive semantics [12] – as a multidimensional cognitive domain consisting of several profiles. Langacker [13] perceives the meaning as a matrix of cognitive domains. The profile and the base of the meaning can be separated in semantic structures by the opposition of a figure and a background through which the sense of a given word is profiled in the language use. According to Langacker, the cognitive domain represents a limited area of a given unit's use. According to Langacker [14], there is no final, exhaustive list of domains; what domain we work with depends on our

particular goal. Bańcerowski's ([9], [11]) cognitive domain can be either a simple or a complex, experience-based notion containing all related knowledge and experience. According to him, each phrase evokes one or more cognitive domains. The profile is the result of profiling where the attention is focused on a concrete element of a particular cognitive base consisting of one or several cognitive profiles whereby this element is accentuated. After studying materials from lexicons, corpuses, religious texts and surveys ([9], [11]), he defines four main domains of the notion of mother: family, value, time and sacrum, assigning them individual phrases.

Bartmiński ([15], [16]) operates with a more detailed classification and attempts to reconstruct general concepts of a good mother through idioms, corpus research, research of folk songs, and surveys. He distinguishes biological, social, psychosocial, psychological, practical, and ethical aspects of the notion of mother. From the biological point of view, it is important that the mother gives birth to one or more children, she breastfeeds her child after the childbirth and the child inherits some of her features. Childcare dominates in the social area. This aspect also covers the mother's position in her family. Upbringing becomes prominent because the mother teaches her children and, when needed, punishes them (or beats them as part of punishment). According to Bartmiński, the mother gives advice to her children, teaches her daughter how to work etc. Obedience and gratitude are expected of children. The psychosocial aspect covers tenderness, warmth, and good heart of a mother who loves her children unconditionally. This aspect creates the space for children's love of their mother. According to this author, the psychological aspect of mother includes other internal characteristics such as sensitivity, understanding, wisdom, and strictness. The practical aspect covers usual roles of a mother who runs a household, gives orders, is overworked, is the only and irreplaceable person for her children and knows her child in every situation. As we can see, the understanding of the concept of mother has been widely affected by individual linguistic and later transdisciplinary tendencies. Through development of linguistic interpretation, progressed from a strictly structuralistic view to cognitive, psychological, and ethnographic approaches, as well as trough stimuli from the field of corpus and discourse analysis, we approached the complexity of the semantics of the word linked to stereotypes and associations in the mental lexicon of users of the particular language. In the next section of the paper, we will seek to capture potential mental processes used in evaluating the image of mother, based on linguistic contexts where the corresponding word is used.

2 THEMATIC AND AXIOLOGICAL ANALYSIS

2.1 Contextual distribution as a basis of axiological interpretation of the word *Mutter* 'mother'

The linguistic image of mother is determined by extralinguistic reality, cultural customs, psychosocial relationships, and expectations based upon them. However, it

is an interdisciplinary question where psychological approaches are particularly significant. In addition to the semantic, distributional, context, and thematic analysis, we have integrated into our research also psychological terms figure, background and profile, originating in the Gestalt psychology, which are used also by cognitive linguists to explain semantic issues. These are terms that Bańczerowski ([9], [10]) uses to describe the concept of mother. However, we will seek to use these terms to describe an axiological process that participants in communication activate in the general use of language. We understand the figure as a lexical meaning of a unit in a collocation. In this case, we perceive the background in connection with an evaluation parameter used as part of the axiological process while using a unit with an evaluation component. At the same time, we develop a scheme according to which *"the evaluation subject applies an axiological function whose field of arguments is exhausted by a class with a single element designated as an evaluation parameter, and the field of values is exhausted by the class of evaluation concepts"* [17, p. 42]. With words oriented at the meaning of the axiological function (talented, genius, diligent, lazy, cowardly…) the point is that the axiological function assigns an implied component *"good"* or *"bad"* [17, p. 43] to the corresponding category (without referring to the value of appearance). By using collocates of the word *Mutter* 'mother' to describe the axiological process, we will seek to submit the evaluation parameter to a finer analysis and link it to the pair of terms profile – background. The profile is linked to the linguistic and extralinguistic implicates which are activated when one uses a lexical unit (and may initiate a positive or negative evaluation response), so they profile the meaning of a given unit. In that case, contextual dependency is characteristic of the profile: what kind of response the stimulus elicits depends on the context or on the information in the background. In this context, we can point out to extensive research of the context effect in the fields of cognitive psychology and marketing which develop analogical assumptions in that particular field and corroborate them with concrete proofs ([18], [19], [20], [21]). When using the word *Mutter* 'mother', the evaluation parameter, to which the axiological function assigns the evaluation concept, depends on the linguistic and extralinguistic context (circumstances of a situation, cultural particularities, individual abilities, and previous experience of the language user etc.). We say that the set of contextual circumstances profiles the meaning of this word (represents its profile) and as part of this profiling it identifies, among others, the evaluation parameter of the axiological function. Because of the binary nature of the opposition figure – background, we propose the following analytical scheme: the argument of the axiological function (the evaluation parameter) is an ordered pair of the figure (a lexical meaning of a lexeme) and the background (the context of evaluation) whereby the evaluation function assigns an evaluation concept to the above-mentioned ordered pair. The field of values of an evaluation function as such can thus be perceived as a Cartesian product of the set of lexical meanings and contexts;

the set of all ordered pairs where the first component is an element of the set of lexical meanings of units evoking axiological notions and the second component originates in the set of contexts in which these words are used and in relation to which it is necessary to relativize the application of the evaluation function. When we look at the second component of the argument more closely, we can divide it into n-ths of various types of contexts (the situation context including the way of expressing a certain characteristic, a cultural context including rules of behaviour in a particular society etc.).

2.2 Thematic analysis

As a background for the thematic analysis of collocations with the base word mother, we have used a complete collocation profile [22] created on the basis of corpus tools ([23], [24]), which contains system and text collocations (compare [25], [26], [27]). In order to define the image of mother more closely in a particular case, we need to know the context. After examining individual context appearances of collocations in the corpus by using the Sketch Engine tool (see examples), we have come to the following thematic subgroups linked to mothers or motherhood in execution of contexts. Following up on the previous study [28], we provide examples of individual thematic subgroups from the above-mentioned collocation profile, using Ďurčo's collocation matrix [29] based on the morphological principle (M. stays for Mutter, m. stays for mother):

I. Body characteristics and their implications:
a) age
Adjective + Noun: *alte M*. 'old m.', *betagte M*. 'elderly m.', *junge M*. 'young m.', *minderjährige M.* 'minor m.'

b) health condition:
Adjective + Noun: *alkoholkranke M.* 'alcoholic m.', *(HIV/...)* - *infizierte M*. 'HIV infected m.', *kranke M*. 'ill m.', *schwerkranke M*. 'severely ill m.', *pflegebedürftige M.* 'm. requiring care'
Noun + Verb: *M. erlitt etw*. *(Schock/Nervenzusammenbruch/Trauma/...)* 'm. suffered from sth. (shock/nervous breakdown/trauma…)'

II. Motherhood:
a) stages of motherhood:
Adjective + Noun: *frischgebackene M*. 'fresh m.', *werdende M.* 'future m.', *potenzielle M.* 'potential m.'; *M. bei der Geburt* 'm. in labour'
Noun + Verb: *M. werden* 'to become a m.'

b) activities:
Adjective + Noun: *einkaufende M.* ‘shopping m.’, *stillende M.* ‘breastfeeding m.’
Noun + Verb: *M. rettete ihr Kind* ‘m. saved her child’, *M. stillt ihr Baby/ihr Kind* ‘m. breastfeeds her child’, *M. trägt ihr Kind auf dem Arm* ‘m. carries her child in arms’, *M. vernachlässigt ihr Kind* ‘m. neglects her child’

c) experience linked to motherhood:
Adjective + Noun: *erfahrene M.* ‘experienced m.’, *frischgebackene M.* ‘new m.’, *kinderreiche M.* ‘m. with many children’, *mehrfache M.* ‘multiple m.’*, x-fache M.* ‘x- times m.’

III. Mother in relationships:
a) the relationship between the mother and her child:
a. 1) the biological relationship between the mother and her child:
Adjective + Noun: *biologische M.* ‘biological m.’, *eigene M.* ‘own m.’, *leibliche M.* ‘own, biological m.’
a. 2) emotional relationship between the mother and her child:
Adjective + Noun: *liebende M.* ‘loving m.’, *geliebte M.* ‘beloved m.’
Adjective + Verb: *M. liebt ihre Kinder* ‘m. loves her children’

b) legal relationships:

b. 1) family law relationships:
Adjective + Noun: *ledige M.* ‘single m.’, *geschiedene M.* ‘divorced m.’, *verheiratete M.* ‘married m.’, *unverheiratete M.* ‘unmarried m.’, *verwitwete M.* ‘widowed m.’

b. 2) criminal law and similar relationships:
Adjective + Noun: *angeklagte M.* ‘accused m.’, *getötete M.* ‘killed/murdered m.’, *verschwundene M.* ‘missing m.’
Noun + Verb: *M. tötete ihr Kind/ihren Sohn/ihre Tochter/ihr Baby* ‘m. killed her child/son/daughter/new-born baby), *M. klagt jdn. an* ‘m. accuses someone’

b. 3) socioeconomic status, labour law relationships, profession:
Adjective + Noun: *erwerbstätige M.* ‘working m.’, *nicht berufstätige M.* ‘unemployed m.’, *nicht erwerbstätige M.* ‘ unemployed m.’, *notleidende M.* ‘needy m.’, *teilzeitbeschäftigte M.* ‘m. in part-time employment’
Noun + Verb: *M. braucht Hilfe* ‘m. needs help’, *M. arbeitet* ‘m. works’

IV. Personality of mother:
a) character:
Adjective + Noun: *aggressive M.* ‘aggressive m.’, *dominante M.* ‘dominant m.’, *gefühllose M.* ‘emotionless m.’, *fleißige M.* ‘diligent m.’, *liebe M.* ‘kind m.’, *liebevolle M.* ‘loving m.’, *nachsichtige M.* ‘tolerant m.’

b) feeling, emotions, mental state:
Adjective + Noun.: *geschockte M*. 'shocked m.', *gestresste M*. 'stressed m.', *glückliche M*. 'happy m.', *verzweifelte M*. 'desperate m.'

c) behaviour, methods and models of upbringing:
Adjective + noun: *aggressive M*. 'aggressive m.', *dominante M*. 'dominant m.', *fürsorgliche M.* 'caring m.', *interessierte M*. 'interested m.', *schimpfende M*. 'swearing m.', *rauchende M.* 'smoking m.'

d) the identity of mother:
Adjective + Noun: *deutsche M*. 'German m.', *gute M*. 'good m.', *moderne M*. 'modern m.', *schlechte M.* 'bad m.'
Noun + Verb: *meine M. kommt aus* ... 'my m. comes from…'

2.3 Process of the collocation base word profiling in context

When we study the appearance of individual collocations in the context, we can pay attention to the process of profiling. We perceive the thematic characteristic (see the previous part) in connection with the contextual information; its identification is linked to the knowledge of appearance of a given collocation in the contexts. The thematic circle of the collocation can be perceived as a domain and the collocation itself as a figure. When the collocation such as an aggressive mother appears in a different domain, a different image is profiled. This fact is evident in multivalent attributes (aggressive mother, dominant mother; the psychological or behavioural characteristic in relation to a child or in general). The profiling is contextually dependent; as an outcome, on the basis of contextually conditioned thematic characteristics, we can demonstrate that by spotlighting the figure through the prism of a type of theme, we identify different evaluation parameters to which the axiological function assigns different evaluation notions. The thematic characteristic – as a contextually conditioned parameter – thus participates in profiling the meaning of the given unit and also in the nature and the outcome of the process of evaluation, e.g., the evaluating subject can in one case evaluate the dominant mother as complying with the evaluation standard, i. e., good, and in another case it can be the opposite. Again, we can point to the significance of the theory of context effect.

When we change the domain, we get a different profile of the figure, e.g., the meaning of collocations such as *Hausfrau und Mutter* 'housewife and mother', *Ehefrau und Mutter* 'wife and mother', *Frau und Mutter* 'woman and mother', *Mutter und Lehrerin* 'mother and teacher' profiles differently depending on the chosen domain. When we project these figures onto backgrounds such as "socioeconomic status, profession" and "social-family relationships", we get different profiles. Depending on the context, it can be the same person who fulfils

various roles in her life, e.g., mother and teacher. In this case, it can also be a positive or negative evaluation of that person because the word teacher can be interpreted on the background of a nice, patient, and creative person who likes children and can lead and motivate them thanks to a well-chosen way of communication but on the background of a strict and authoritative person, it can lead to negative expectations. Another option is that they are two different persons who are the subjects of the same sentence. In this case, they are often participants in a conflict. Collocations such as *liebe Mutter* 'dear mother', *liebevolle Mutter* 'loving mother', *nachsichtige Mutter* 'tolerant mother', *aggressive Mutter* 'aggressive mother', *schimpfende Mutter* 'swearing mother' can be interpreted on the background of themes of behaviour and habits; mental states and internal characteristics of the mother, or a relationship between the mother and the child. Depending on the chosen domain, the speaker automatically operates with specific, empirically sound psychological, social and other expectations, stereotypes and implications of meanings that determine his evaluation of the profiled entity. As a consequence, the result on the good-bad axis is strongly affected by extralinguistic factors including beliefs and experience of the language user. In this context we can remind ourselves of Miko's earlier conclusions about anchoring of a language expression in the extralinguistic context and involvement of an experience-related complex in the process of its specification [30, p. 14.], however, they were not formulated in relation to the process of interpretation but production. Such perspective opens a wide range for interdisciplinary research dominated by psychological[1] and pragmatic approaches.

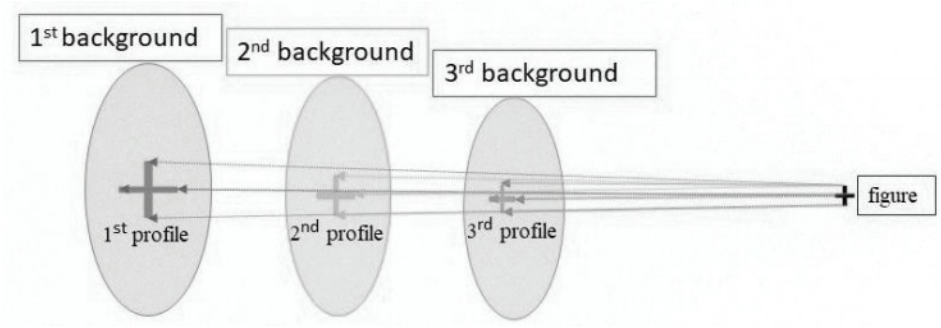


Fig. 1. Simplified profiling of a figure on the basis of various backgrounds

[1] For pragmatic aspects of collocations in spoken communication, see [31].

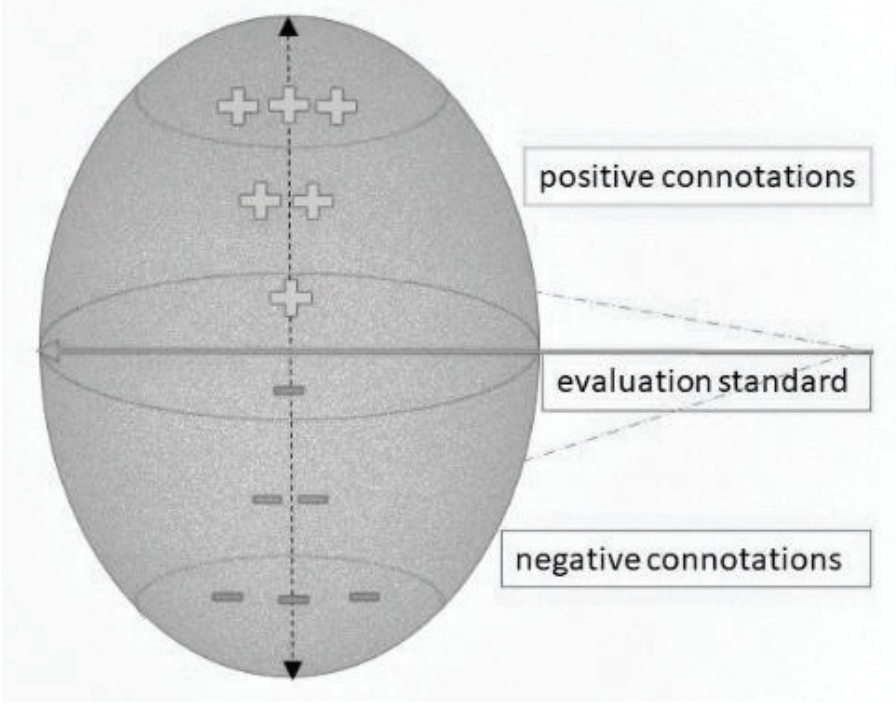


Fig. 2. Simplified scheme of evaluation standard and connotations

Collective experience, as written about by Kmecová [7, p. 70], plays an important role as part of understanding the linguistic image of evaluation. The evaluation standards in the sense of standardised evaluation parameters, and the stereotypical attribution of values on a given value scale, "are in a certain sense collectively conventionalised mental representations of attributes, which are in the base of the evaluation of the entities, as well as stereotypical attribution of values on the basis of these attributes" [32, p. 11].

Love, care, and proper upbringing are important concepts in the research of Rakhimova, Mukhamadiarova and Tarasova [33], who analysed German proverbs describing the family and concluded that *"the relationship between mother and child is characterized, in the first place, by the indispensability of maternal love and care. ... Proverbs emphasis that the main task of parents is to educate children properly and prepare them for adulthood"* [33, p. 1054]. These factors seem to form the basis of the mother's assessment standard in German language culture. Based on how the mother performs her tasks as defined above, it is decided whether she is a good mother or a bad one. Thus love, care, and proper upbringing can be considered the basis of axiological Qualia in the sense as Hanks [34] writes.

3 CONCLUSION

Just as with other polysemantic lexical units, the meaning of the German word *Mutter* 'mother' is specified by the context in which it appears. As a lot of research by earlier scholars showed, substantive denominations of "mother" are connected with the notion that is – from the point of view of cognitive linguistics – characterised by a high degree of complexity. This has been proved also by results of the distribution and thematic analyses of these words on the basis of corpus material. Processing of contextual information (collocations and wider contextual data) allows us to reconstruct prototypical characteristics of "mother" and the related concepts of German speakers. In the subsequent analysis focused on the axiological aspects of the notion of mother, employment of psychological terms figure, background, and profile as well as the theoretical bases linked with the theory of context effect has proved to be beneficial. (Further research may in the future relate to the relationships of the mentioned concept with the phenomena of semantic prosody and semantic preference (for example [35], [36]). The submitted proposal is that the argument of the axiological function (the evaluation parameter) is an ordered pair of the figure (a lexical meaning of a lexeme) and the background (the context of evaluation), whereby the evaluation function assigns an evaluation concept to the above-mentioned ordered pair. Similarly, the thematic characteristic of words *Mutter* 'mother', derived from the thematic analysis of corresponding contexts, participates in profiling the meaning of the given unit and thus also in the nature and result of the evaluation process. The application of this process onto the context of words *Mutter* 'mother' recorded in the corpus material allows us to outline the unconscious axiological processes in evaluating the image of mother by German speakers.

References

[1] Bierwisch, M. (1969). Strukturelle Semantik. Deutsch als Fremdsprache. Zeitschrift für Theorie und Praxis des Deutschunterrichts für Ausländer, 6(2), pages 66–74.

[2] Lakoff, G. (1987). Women, Fire, and Dangerous Things. What Categories Reveal about the Mind. Chicago and London, The University of Chicago Press, 634 p.

[3] Lakoff, G. (2016). Moral Politics. How Liberals and Conservativs Think. Chicago, The University of Chicago Press, 512 p.

[4] Kiefer, F. (2007). Jelentéselmélet. Budapest, Corvin/MTA Nyelvtudományi Intézet, 381 p.

[5] Wierzbicka, A. (1996). Semantics. Primes and Universals. New York, Oxford University Press, 512 p.

[6] Orgoňová, O., and Bohunická, A. (2012). Lexikológia slovenčiny. Praha, COLUMBUS, 282 p.

[7] Kmecová, S. (2016). Pec sa im pováľala ... materstvo a jeho obraz v slovenskej a slovinskej frazeológií. Slovanski jeziki na stičišču kultur. Mednarodno znanstveno srečanje mladih humanistov. Ljubljana, National and University Library, pages 61–74.

[8] Černá, Z., and Čech, R. (2019). Analysis of the Lemma mateřství (Motherhood). Jazykovedný časopis, 70(2), pages 244–253.
[9] Bańczerowski, J. (2006). Az anya fogalmának nyelvi képe a magyar nyelvben. Magyar Nyelvőr, 130(3), pages 261–271.
[10] Bańczerowski, J. (2007). A család fogalma a világ magyar nyelvi képében (egy kérdőíves anyag tükrében). Magyar Nyelvőr, 131(2), pages 235–245.
[11] Bańczerowski, J. (2008). A világ nyelvi képe. Budapest, Tankönyvkiadó, 356 p.
[12] Langacker, R. (1995). Wyklady z gramatiky kognitywnej. Lublin, Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, 186 p.
[13] Langacker, R. (1986). An Introduction to Cognitive Grammar. Cognitive Science A Multisciplinary Journal, 10(1), pages 1–40.
[14] Langacker, R. (2008). Cognitive Grammar. A Basic Introduction. Oxford, University Press, 562 p.
[15] Bartmiński, J. (2008). Polski stereotyp matki. Postscriptum Polonistyczne, 1(1), pages 33–53.
[16] Bartmiński, J. (2016). Polský stereotyp matky. Jazyk v kontextu kultury. Praha, Karolinum, 168 p.
[17] Dolník, J. (2007). Lexikológia. Bratislava, Univerzita Komenského, 293 p.
[18] Cavanagh, P. (1991). What´s up in top-down processing? Representations of Vision: Trends and tacit assumptions in vision research, pages 295–304.
[19] Deaton, A., and Stone, A. A. (2016). Understanding context effects for a measure of life evaluation: how responses matter. Oxford Economic Papers, 68(4), pages 861–870.
[20] Smyth, D. J., Dillman, D. A., and Christian, L. M. (2007). Context effects in Internet surveys: new issues and evidence. Oxford Handbook of Internet Psychology, New York, Oxford University Press, pages 429–446.
[21] Trueblood, J. S. et al. (2013). Not just for consumers: Context effects are fundamental to decision-making. Psychological Science, 24(6), pages 901–908.
[22] Accessible at: http://www.vronk.net/wicol/index.php/Mutter.
[23] Ďurčo, P. (2008a). Zásady spracovania slovníka kolokácií slovenského jazyka. Accessible at: http://www.vronk.net/wicol/images/Zasady.pdf.
[24] Ďurčo, P. (2008b). Konzept eines zweisprachigen Kollokationswörterbuchs. Prinzipien der Erstellung Am Beispiel Deutsch ↔ Slowakisch. Accessible at: http://www.vronk.net/wicol/images/Kollokationen_Durco_SK.pdf.
[25] Čermák, F. (2010). Lexikon a sémantika. Praha, NLN, 357 p.
[26] Vajičková, M. (2017). Theoretische Aspekte der Kollokationen. In Kollokationen im Sprachsystem und Sprachgebrauch. Ein Lehrbuch, pages 12–50, Nümbrecht, Kirsch Verlag.
[27] Vajičková, M. (2019). Textlinguistische Aspekte der Kollokationen. In Kollokationen im Sprachsystem und Sprachgebrauch. Ein Lehrbuch, pages 133–162, Nümbrecht, Kirsch Verlag.
[28] Braxatorisová, A. (2020). Prístupy k významu slova matka a jeho tematické kontexty v nemeckom jazyku. Philologia, 30(2), pages 75–87.
[29] Ďurčo, P. (2019). Ansätze zur Analyse der Kollokationen. In Kollokationen im Sprachsystem und Sprachgebrauch. Ein Lehrbuch, pages 51–132, Nümbrecht, Kirsch Verlag.
[30] Miko, F. (1970). Text a štýl. Bratislava, Smena, 167 p.
[31] Tomášková, S. (2019). Pragmatische Aspekte der Kollokationen in mündlicher Kommunikation. In Kollokationen im Sprachsystem und Sprachgebrauch. Ein Lehrbuch, pages 163–190, Nümbrecht, Kirsch Verlag.

[32] Dolník, J. (2005). K otázke odrazu hodnôt v jazyku. Jazykovedný časopis, 56(1), pages 3–12.
[33] Rakhimova, A. E., Mukhamadiarova, A. F., and Tarasova F. K. (2019). Linguistic and Cultural Characteristics of Proverbs Describing Family Relations in the German Linguistic Picture of the World. Humanities & Social Sciences Reviews, 7(6), pages 1048–1055.
[34] Hanks, P. (2010). Wie man aus Wörtern Bedeutungen macht: Semantische Typen treffen Valenzen. In Sprachliches Wissen zwischen Lexikon und Grammatik. Jahrbuch Institut für Deutsche SpracheWalter de Gruyter, pages 483–503.
[35] Oster, U., and Lawick, H. (2008). Semantic Preference and Semantic Prosody: A Corpus-Based Analysis of Translation-Relevant Aspects of the Meaning of Phraseological Units. Translation and Meaning, 8, pages 333–344.
[36] Busse, D. (2012). Frame Semantik: Ein Kompendium. Berlin/Boston: Walter de Gruyter, 888 p.

sciendo
DOI 10.2478/jazcas-2021-0059

CZECH TRANSLATIONS OF THE GOSPEL OF MATTHEW FROM THE DIACHRONIC POINT OF VIEW – PLUS ÇA CHANGE…

RADEK ČECH[1] – JÁN MAČUTEK[2,3]
– PAVEL KOSEK[4]
[1] Department of Czech Language, University of Ostrava, Ostrava, Czech Republic
[2] Department of Mathematics, Constantine the Philosopher University, Nitra, Slovakia
[3] Mathematical Institute, Slovak Academy of Sciences, Bratislava, Slovakia
[4] Masaryk University, Brno, Czech Republic

ČECH, Radek – MAČUTEK, Ján – KOSEK, Pavel: Czech translations of the Gospel of Matthew from the diachronic point of view – Plus ça change… Journal of Linguistics, 2021, Vol. 72, No 2, pp. 656 – 666.

Abstract: The paper focuses on dynamics of changes of several linguistic and text properties in diachronic development of Czech. Specifically, we analyze the proportion of identical word-forms (types), the average type length, text length, the proportion of hapax legomena, the moving average type-token ratio, and entropy. For the analysis, seven translations of the Gospel of Matthew from the 14th to the 21st century were used. The study reveals some differences in dynamics of changes of particular properties.

Keywords: Gospel of Matthew, diachronic development, dynamics of changes, properties of text, Czech language

1 INTRODUCTION

Language is a phenomenon undergoing continuous changes. It is well known that there are periods of more intense changes as well as periods of a relative stability of the language system. Further, in particular linguistic planes, one can find differences in the dynamics of changes. For instance, diachronic development of a lexical plane seems to be more or less a continual process, where more dramatic changes, if they appear at all, are usually caused by extra-linguistic factors (such as the National Revival in the 19th century in Czech). On the other hand, in diachronic development of both phonological and morphological planes, we can observe periods of very intense changes followed by periods in which these properties of language remain almost constant.

Up to now, historical linguistics has brought plenty of analyses describing and explaining the development of phonological, grammatical, and lexical properties of language. Following this direction of linguistic research, in this study, we focus on the historical development of some lexical properties and text characteristics which, so far, have not been in the center of attention of historical linguists, at least for

Czech. Our aim is to observe, compare, and explain the dynamics of changes in these properties. For the analysis, we use seven different translations of a single text, namely, of the Gospel of Matthew, which were published from the 14th to the 21st century. This approach allows us to control to a certain extent the impact of factors, such as genre, topic, authorship etc. on the observed phenomena. If we use texts of the same genre, topic etc. (in our case different variants of a single text), we know that potential differences in results are not caused by these factors. For instance, it is well known that the average word length varies, depending on the genre and topic. So, if we want to observe potential changes in word length during historical development of a language, it is difficult to find samples of texts which share the same characteristics in particular periods. Thus, the choice of different variants of a single text seems to be one of acceptable approaches, although of course not the ideal one, cf. [1].

As we mentioned above, dynamics of different language (or text) properties varies. For illustration, let us start with a comparison of the verse 2.1 from the oldest (1) and the youngest (2) Czech translation of the gospel.

(1) *Protož když* ***sě*** *jest* ***urodil*** *Ježíš v* ***Bethlémi*** *židovském za* ***dnóv*** *krále Heroda, tehdy mudráci* ***ode vzchoda*** *sluncě přišli sú do Jeruzaléma*.
(*Bible of Dresden*, 1370s)

(2) *Když* ***se*** *Ježíš* ***narodil*** *v judském* ***Betlémě*** *za* ***dnů*** *krále Heroda, hle, mágové* ***od východu*** *přišli do Jeruzaléma.*
(*Czech Study Translation*, 2009)

‘Now when Jesus was born in Bethlehem of Judaea in the days of Herod the king, behold, there came wise men from the east to Jerusalem.’[1]

At first sight, differences are evident. First, the length of the verse (measured in the number of tokens) differs – 22 tokens in (1) versus 18 tokens in (2). Second, the words that occur in both texts but differ in their form (e.g., *urodil – narodil* ‘was born’) are highlighted in bold. Third, there are only nine identical word forms which appear in both (1) and (2), namely: *když* ‘when’, *Ježíš* ‘Jesus’, *v* ‘in’, *za* ‘during’, *krále* ‘king’, *Heroda* ‘Herod’, *přišli* ‘came’, *do* ‘into’, *Jeruzaléma* ‘Jerusalem’. However, there are some properties of language and text which are not so “visible” at first glance, for instance, word length or lexical diversity. Again, for illustration, the average type length in (1) is 4.72 characters, while in (2) it is 5 characters. As for lexical diversity (or vocabulary richness), if we compare the type-token ratio (*TTR*) of the second chapter of the gospel in the oldest and youngest (2) Czech translation,

[1] This is the modern translation of the excerpt, taken from [2].

we get the values of 0.542 and 0.591, respectively. These differences seem to be rather small, especially if one compares them with the number of lexical or word-form changes (see above). So, the question is, whether they follow the variability we observed above or whether they represent more stable properties.

In this study, we start with a comparison of the vocabulary of particular text variants where the differences are most obvious. Then, we focus on both word characteristics (length of word types) and some properties of text (text length and several measures of lexical diversity). As for the analysis of word-type length, our goal is to find out if the relatively high number of phonological, morphological, and lexical changes which have taken place in Czech since the 14th century is followed by changes of this word property. If so, our aim is to observe its dynamics and to compare it with the dynamics of changes in vocabulary. Text characteristics (as opposed to characteristics of language) and their potential changes can be seen from a somewhat broader perspective. They can be explained as a result of a combination of changes in the language system as well as of different translation techniques.

2 LANGUAGE MATERIAL AND METHODOLOGY

For the analysis, we use translations of the Gospel of Mathew from the 14th to the 21st century which were part of

- Bible drážďanská (BiblDrážď), Bible of Dresden (the 70s of the 14th century),
- Bible olomoucká (BiblOl), Bible of Olomouc (1417),
- Bible Melantrichova (BilMel), Melantrich's Bible (1570),
- Bible kralická (BiblKral), Bible of Kralice (1596),
- Bible svatováclavská (BiblSvat), Bible of St. Wenceslas (1677),
- Nový zákon (BiblSýk), New Testament (1909),
- Český studijní překlad Bible (ČSP – Bible), Czech Study Translation (2009).

From the great number of translations of this gospel, we tried to select the ones which are considered significant in terms of development of the Czech biblical translation and which well represent the individual periods ([3], [4]). We decided not to use the first chapter of the Gospel of Mathew because of its specific nature. It introduces the genealogy of Jesus and it mainly consists of the list of names.

The following properties were used for comparison of the texts: 1) the proportion of identical word-forms (types) in pairs of texts (*PIV*), 2) the average type length (*AVL*), 3) text length (*N*), 4) the proportion of hapax legomena (*PHL*), 5) the moving average type-token ratio (*MATTR*), 5) entropy (*H*).

PIV is determined by the number of identical word-forms which occur in two texts. It is calculated as

$$PIV = \frac{|V_i \cap V_j|}{|V_i|} ,$$

where V is the list of word-forms of the text (i.e., list of types) and i and j are texts. In this study, we calculate *PIV* in relation to the BiblDrážď only (i.e, in all calculations V_i is the set of word-forms which occur in the Gospel of Matthes from BiblDrazd). It allows us to interpret *PIV* as the proportion of changes relative to the oldest translation which is considered as a "reference point" here.

AVL is defined as the arithmetic mean of type lengths in a text. The type length is calculated as the number of characters (e.g., the word *východu* is 7 characters long).

N is determined by the total count of all tokens in a text. A token is a graphic word, i.e., a sequence of characters separated by spaces.

The last three indices are associated with the lexical diversity of the text. Since it is a relatively complex phenomenon which can be seen (and measured) from several points of view, we decided to apply the following methods. *PHL* is the proportion of hapax legomena in the text. *MATTR* is based on the segmentation of text into overlapping chunks (so-called windows), where the type-token ratio is calculated for each window. It is determined as the arithmetic mean of the type-token ratios in all windows, i.e.,

$$MATTR = \frac{\sum_{i=1}^{N-L} V_i}{L(N-L+1)} ,$$

where N is the number of tokens in the text, L is the length of the window, V_i is the number of types in the window. In this study we use $L = 100$. H is usually interpreted as a measure of system uncertainty. In the case of a text, it expresses the degree of lexical and word form diversity. Specifically, the greater the value of entropy, the more diversified (i.e., less concentrated) the vocabulary is. It is calculated as

$$H = \log_2 N - \frac{1}{N} \sum_{r=1}^{V} f_r \log_2 f_r ,$$

where N is the frequency of tokens in the text and f_r is the frequency of a given token.

For the text processing and computation, we used the QuitaUp [5] (for *N*, *PHL*, *MATTR*, *H*) and R-software [6] (for *PIV*, *AVL*).

3 RESULTS

Let us start with the comparison of *PIV,* which should be the most distinct indicator of changes (cf. Section 1). In Table 1 and Figure 1, we can see the

proportions of word forms (types) which BiblOl, BiblMel, BiblKral. BiblSvat, BiblSýk, and ČSP share with BiblDrážď. As we mention in Section 2, we consider BiblDrážď the reference point since it is the oldest surviving Czech translation. The higher the value of *PIV*, the more similar the texts.

Bible	***IWF***	***PIV***
BiblOl	2,299	0.530
BiblMel	1,525	0.351
BiblKral	1,464	0.337
BiblSvat	1,467	0.338
BiblSýk	1,375	0.317
ČSP	1,116	0.257

Tab. 1. Proportions of word forms (types) (*PIV*) which are identical in given texts and the BiblDrážď (which contains 4340 types). *IWF* is a total number of word forms occurring in both BiblDrážď and the given translation

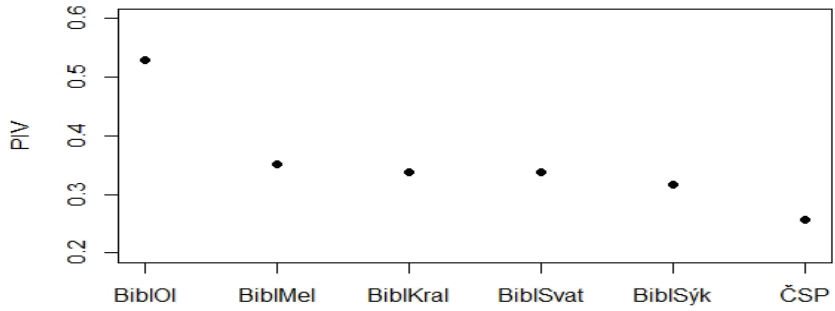


Fig. 1. Proportion of identical word forms (types) (*PIV*) in given texts and the BiblDrážď

The results show that the number of differences is rapidly increasing (i.e., the value of *PVI* is decreasing) from BiblDrážď to BiblMel. The difference between BiblDrážď and BiblOl corresponds with [3, pp. 47, 53] who claim that these two translations were created independently of each other.[2] As for BiblMel, it is considered a translation independent from the older ones. Moreover, there is a 200-year time span between BiblMel and BiblDrážď and more than 150 years between BiblMel and BiblOl, which also can be an important factor. Last but not least, the fundamental phonological, morphological, and syntactic changes which took place in Czech in this period (i.e., from the 14th to the 16th century) influenced the form of BiblMel substantially. By contrast, the next translations up to the beginning of the 20th century

[2] BiblDrážď and BiblOl are copies of the same (lost) source text of the original text of the Old Czech Bible of the 1st edition from the 60s of the 14th century. However, BiblOl differs in the translation of the Gospel of Matthew known as Gospel of Matthew with homilies (from the 70s of the 14th century), cf. [3, pp. 47, 53].

are very similar (in the sense of the method we used). This finding is in accordance with the textual relations between these three translations that are known from secondary sources [3, pp. 180, 212]. The translators/editors of the Gospel of Matthew from BiblKral were influenced by the version published in BiblMel. The translators/editors of BiblSvat were influenced by both BiblMel and BiblKral. BiblSvat, respected by the Catholic Bible translators/editors, was the dominant Catholic translation until the 20th century [4, p. 1887]. All these translations can be interpreted as a result of a specific translation tradition. The "power" of this translation tradition is particularly evident in the fact that each of the BiblMel, BiblKral, and BiblSvat comes from different Czech churches (the Utraquist, Unity of the Brethren, Catholic, respectively). Another factor that has an impact on the language form of the translations is the nature of written Czech. Specifically, a minimum of fundamental changes appear at the phonological and morphological level from the end of the 16th century to the 19th century. Finally, the youngest translation is less similar in comparison to the oldest one. This finding is in accordance with the declared effort of the translators to escape the above mentioned strong translation tradition and to use the language that is commonly used at the beginning of the 21st century. To, sum up, the dynamics of *PIV* development corresponds with current knowledge regarding both the language development and the translation tradition.

Next, let us concentrate on the development of *AVL* (see Table 2 and Figure 2). In comparison to *PIV*, we get a completely different picture. Surprisingly enough, not linguistically important changes of *AVL* took place in the period of more than 600 years. The differences are strikingly small – the biggest difference is the difference of two-tenths of a letter, on average. To get a more complex view of the nature of the data, standard deviations (*SD*) of *AVL* were computed, too. In our case, *SD* expresses a variability of word lengths in particular texts. Again, a great stability of values of *SD* appears throughout the period under investigation. These findings are very unexpected, especially if one realizes both a number of linguistic changes which took place since the end of the 13th century in Czech and diversity of translation strategies. In our analysis, *AVL* seems to be an extremely stable property that resists any development.

Bible	***AVL***	***SD* (*AVL*)**	***N***
BiblDrážď	6.41	2.1	17,327
BiblOl	6.34	2.1	17,818
BiblMel	6.48	2.14	17,548
BiblKral	6.52	2.19	17,621
BiblSvat	6.48	2.15	17,083
BiblSýk	6.52	2.18	17,093
ČSP	6.52	2.16	17,109

Tab. 2. The average type length (*AVL*), the standard deviation of *AVL* (*SD*), and the length (*N*) of individual translations

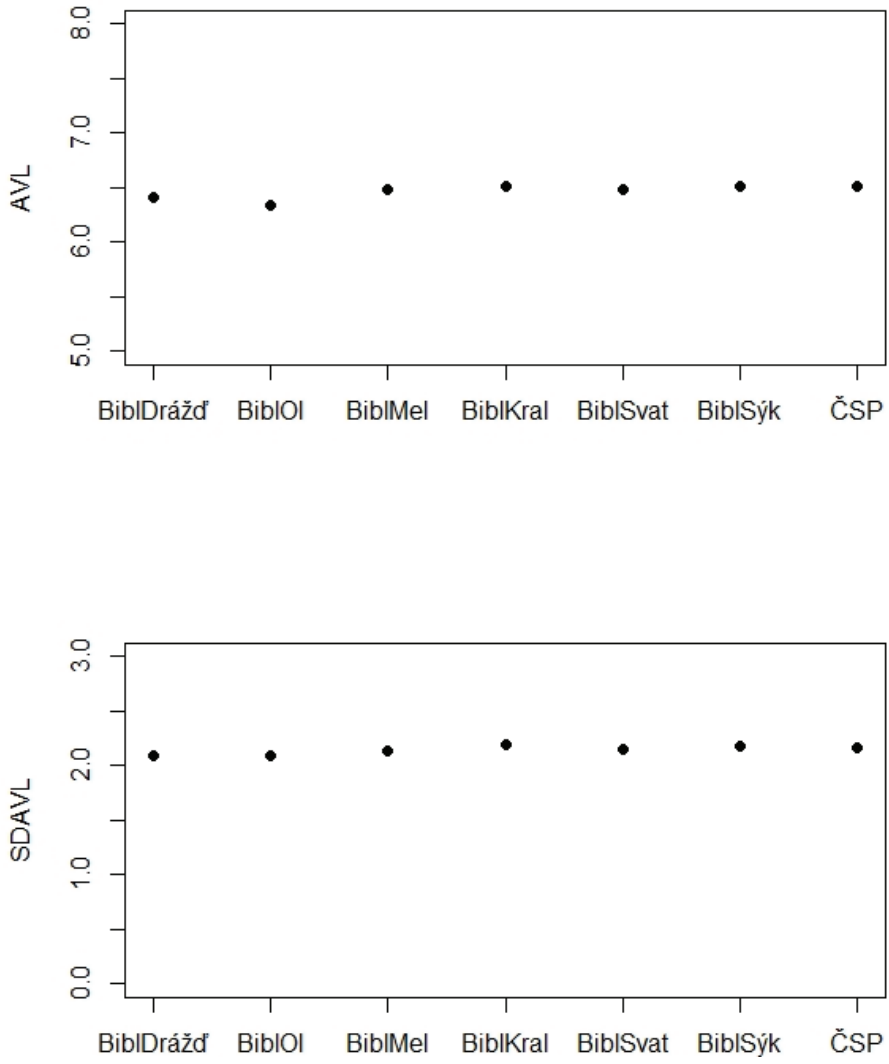


Fig. 2. The average type length (*AVL*) and the standard deviation of *AVL* (*SD*) in individual translations

There are no theoretical reasons to expect any direction in the historical development of the length (*N*) of particular translations, in our opinion. Of course, a variability of *N* can appear, as even a comparison of very short excerpts exemplifies, see (1) and (2) in Section 1. The data presented in Table 2 and Figure 3 reveal no single tendency in historical development. The biggest difference (between BiblOl and BiblSvat) is 735 tokens which means that the length of both texts differs by

approximately 4%. Further, there is an obvious difference between the two groups of texts. Specifically, a variability of *N* is evident among the first four translations, while in the group of the last three texts there are minimal differences of *N*. It can be explained as follows. In the 16th century, humanist scholars discussed the form of original Biblical texts which were used for translation. They were motivated by an effort to eliminate non-original parts which were added to the original during the Middle Ages. Since the end of the 16th century, there has been a relatively high consensus on the form of original Biblical text used for translation. Thus, the minimal variability of *N* in the last three translations in our sample can be interpreted as a consequence of this fact.

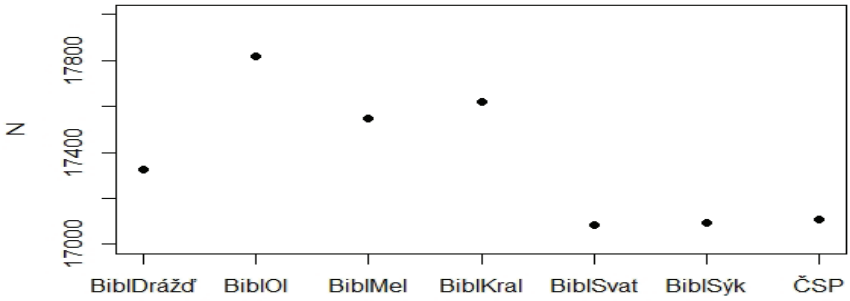


Fig. 3. The length (*N*) of individual texts

Bible	***PHL***	***MATTR***	***H***
BiblDrážď	0.154	0.781	9.981
BiblOl	0.133	0.766	9.826
BiblMel	0.139	0.775	9.896
BiblKral	0.141	0.781	9.936
BiblSvat	0.140	0.781	9.932
BiblSýk	0.144	0.784	9.955
ČSP	0.140	0.779	9.909

Tab. 3. The proportion of hapax legomena (*PHL*), the moving average type-token ratio (*MATTR*), and the entropy (*H*) of individual translations

The measurement of lexical diversity (sometimes called also vocabulary richness) captures how the writer (in our case the translator or translators) "manipulate" the vocabulary at his or her disposal to express the content. Obviously, there are several ways how to perform it. For instance, the usage of synonyms increases the lexical diversity, while a repetition of words (which one can use intentionally to ensure a high text cohesion) decreases it. There is no single method of measuring this property of the text. Here, we use three of them (*PHL*,

MATTR, *H*) to eliminate potential biases of particular methods. The results show (see Table 3 and Figure 4) that there are minimal differences among texts, regardless of the method.

At first sight, these results are rather surprising. Particular translations come from different periods, they were translated from different original texts (even from different languages), published by various churches for different recipients, the translators followed different translation strategies etc. Despite all these circumstances, the lexical diversity of all texts is almost identical. Most likely, there must be some dominant factor that eliminates the expected variability. In our opinion, the sacred nature of the gospel plays a decisive role here. Consequently, the translators strictly adhere to the lexical diversity of the original[3] and they do not dare to be creative in this regard.

4 CONCLUSION

In this study, we analyzed the historical development of several language and text properties of the Czech translations of the Gospel of Matthew to get a picture of their dynamics. We started with the most obvious one, the difference in word forms (*PIV*) in individual texts, and found out that the results are consistent with what is reported in the secondary sources. Further, the differences we detected in the text length (*N*) of particular translations can be explained if one realizes the historical context that influenced the translation process. The analysis of type length (*AVL*) brings the most surprising result we did not expect. This property seems to be extremely stable. Of course, this conclusion is based on the analysis of a very limited sample and we are aware that a much larger and diverse sample has to be examined to get a more reliable picture. Finally, the lexical diversity of individual texts is surprisingly stable in all texts. Here, we suggested the explanation which, however, must be confronted with other analyses of non-sacred texts.

Needless to say, the study is just a first step and have to be considered as a pilot study only. Only further research can corroborate (or falsify) the presented conclusions.

ACKNOWLEDGEMENTS

The study was supported by the project MUNI/FF-DEAN/1556/2019 "Vývoj pronominálních enklitik 'mi', 'ho', 'mu' ve starších českých biblích/The Development of the Czech Pronominal Enclitics 'mi', 'ho', 'mu' in Older Czech Bibles" (Pavel Kosek) and VEGA 2/0096/21 (Ján Mačutek).

[3] There is an open question if there are differences of lexical diversity in particular versions of original texts.

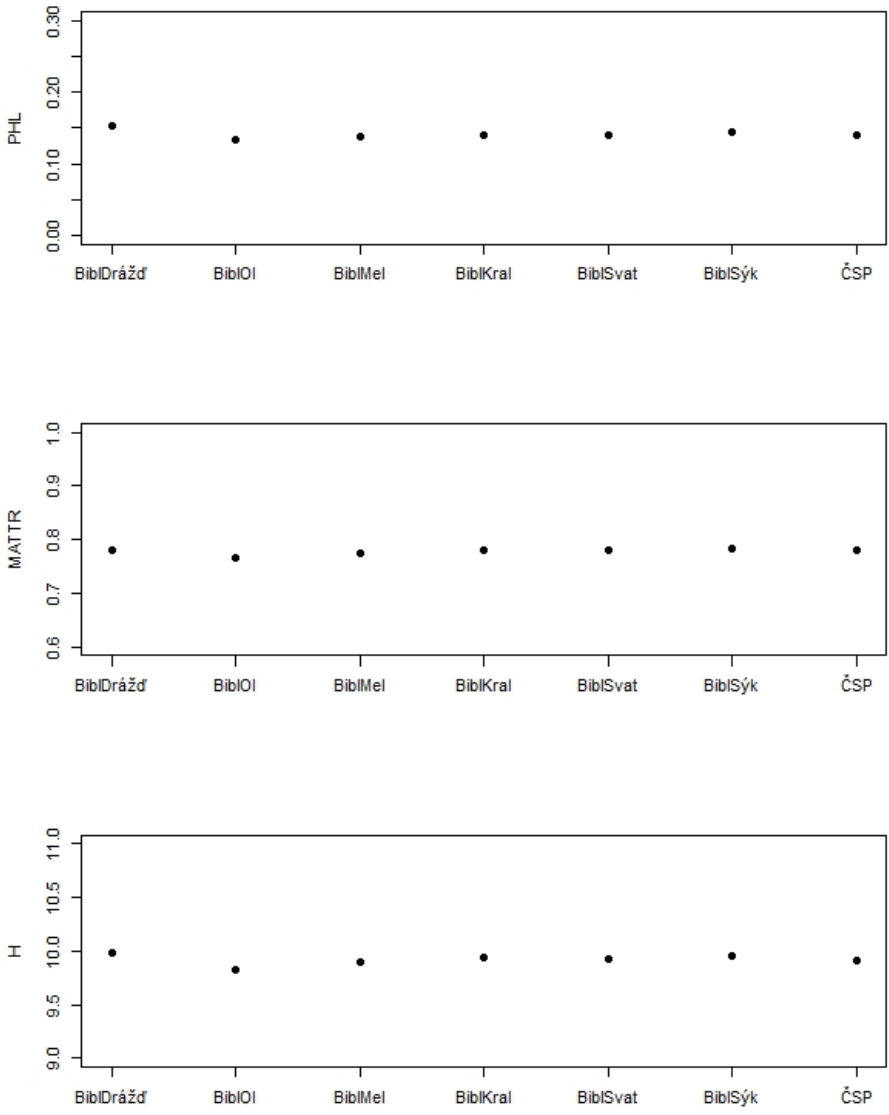


Fig. 4. The proportion of hapax legomena (*PHL*), the moving average type-token ratio (*MATTR*), and the entropy (N) of individual translations

References

[1] Čech, R., Kosek, P., Mačutek, J., and Navrátilová, O. (2020). Proč (někdy) nemíchat texty aneb Text jako výchozí jednotka lingvistické analýzy. Naše řeč, 103, pages 24–36.

[2] King James Bible Online. (2021). Available at: https://www.kingjamesbibleonline.org/.
[3] Kyas, V. (1997). Česká Bible v dějinách národního písemnictví. Praha: Vyšehrad.
[4] Vintr, J. (2008). Bible (staroslověnský překlad, české překlady). In L. Merhaut et al. (eds.), Lexikon české literatury. 4/II U–Ž, Dodatky A–Ř. Praha: Academia, pages 1882–1887.
[5] Cvrček, V., Čech, R., and Kubát, M. (2020). QuitaUp – a tool for quantitative stylometric analysis. Czech National Corpus and University of Ostrava. Available at: https://korpus.cz/quitaup/.
[6] R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/.

sciendo
DOI 10.2478/jazcas-2021-0060

INCOME, NATIONALITY AND SUBJECTIVITY IN MEDIA TEXT

IRENE ELMEROT
Department of Slavic & Baltic Studies, Finnish, Dutch & German, Stockholm University, Stockholm, Sweden

ELMEROT, Irene: Income, nationality and subjectivity in media text. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 667 – 678.

Abstract: This article takes a bird's eye view of how positive or negative sentiments in the news press about countries and nationality nouns seem to reflect the country's general income groups. The study focuses on the four income groups classified by the World Bank and their co-occurrence with positively and negatively classified adjectives from the Subjectivity Lexicon for Czech. A search in the journalistic subcorpus of the SYN series, release 8 of the Czech National Corpus, results in a time line covering three decades. Previous research on subjectivity has either focused on other parts of the Subjectivity Lexicon or on fewer adjectives from other languages. In this article, it is argued that the income groups are treated in descending order, i.e., the higher the income, the more positive the sentiment. Even when the most influential groups in the top and bottom are removed, the result holds. Discourse concerning global war and peace, and the security of different nations, is also detected as a result.

Keywords: income groups, news press, sentiment, nationality, corpus linguistics, Czech language

1 INTRODUCTION

The focus of this article is the overall representation of different economic groups of countries and nationals in the Czech news press over three decades. The study concerns linguistic othering through the positive or negative value of adjectives co-occurring with nouns for these countries and nationals. In this study the "others" are considered to be part of a different income group from Czechs on an average, national level, whereas the same income group as that of the Czech Republic is considered to be the "ingroup". Discourses from the "Western" side of the Iron Curtain entered the Czech Republic at the end of 1989, after the transition from a Communist one-party to a multi-party state. The printed news media was then, and for many years to come, important in providing some kind of relation between the inhabitants, on the one hand, and what Fairclough [1] calls this new "outside", on the other. Many of the former Soviet-led countries are neighbours of the Czech Republic, and even today receive more news coverage due to their geographical location and cultural proximity ([2], [3]) than they do in news articles written in countries farther away. Most of these countries are now economically stronger than before 1989 and,

like the Czech Republic, are classified by the World Bank as High Income countries, with the exceptions of Bosnia and Herzegovina, Bulgaria, Kosovo, Montenegro, Romania, and the Russian Federation, which are all classified as Upper-middle Income countries [4]. In this article, discourse is seen as recurring, prominent claims about something (or someone), that makes a difference in the way the receivers of the discourse talk or write about these things or (groups of) persons [5].[1]

The purpose of this article is to examine whether theories proposed for other languages, such as those of Fairclough [1, p. 29] (see above) and Wodak [5, p. 10] on discursive constructions of national identities (see below), and Chovanec and Molek-Kozakowska [8] on the fluctuation of othering over time, also hold for the printed news press in the Czech Republic. The time span is 1990–2018, covering almost 30 years of democratic governance since the former Communist one-party state. The research aim is to conduct a corpus-based analysis of linguistic othering over time, using a search in which pre-defined adjectives from the Czech Subjectivity Lexicon [9] that co-occur with pre-defined nouns (see below under Material) are collected into a dataset extracted from the corpus. The dataset includes some variables from the corpus, such as publication year and title of the source, but also includes variables using World Bank income classifications.

1.1 Previous research and the contribution of this study

Veselovská describes studies that combine sentiment and discourse in extensive detail [10] and which essentially explain how language is not only a construction of human interaction but an influencer thereof. The sentiment analysis in the Czech Subjectivity Lexicon by Šindlerová et al. [11] is mainly concerned with evaluative verbs, which is one reason for this study to focus on adjectives. Paradis et al. [12] study 42 antonymic adjective pairs in English, focusing on their semantic profiles, whereas this study considers 773 adjectives co-occurring with the specified nouns.

The four income groups of the world's economies, as defined by the World Bank, are rarely used in linguistic studies, but more often in other fields, such as United Nation reports [13] or medical articles ([14], [15]). Language issues relating to these groups or countries have, however, been studied ([16], [17]). Theoretically, this article follows Fairclough's note that there is a need for condensation when researching relations in politics and economics, simply because those subjects are complex in themselves [1, p. 18]. Another of the ideas behind this study is to examine a larger whole (i.e., a bird's eye view over three decades) to see which details emerge, since they may show us the prominent part(s) of the rhetoric [ibid., p. 18] used about these groups over time. It also aims to connect to the ideas in Wodak et

[1] One example is how a male-dominated discourse has meant that many languages have male-only connotations of certain professions [6]; another is how a European-dominated discourse made people believe that all Africans were lazy [7].

al. [5, p. 10] regarding connotations of national identities. The co-occurring adjectives and nouns in this study show how the countries (and their citizens) in these income groups are, and have been, reflected in Czech printed news media over three decades. The binary classification reflected in (positive) peaks and (negative) troughs for the four groups is based on the recently mentioned Subjectivity Lexicon.

2 RESEARCH QUESTIONS AND MATERIAL

The research question of the study is:

In what way does the binary sentiment of the adjectives correlate with the income group?

The hypothesis for this study is that the higher the income group, the more positive their overall level. To demonstrate this, the binary classification of adjectives is represented in the form of graphs.

2.1 Material

Two main sources have been used to create the dataset: the adjectives in the Subjectivity Lexicon, and a journalistic subcorpus of the SYN series, release 8, containing 4,499,370,372 tokens (running words, excluding punctuation) from the Czech National Corpus (see [18]). These adjectives[2] in the Lexicon are classified into one of three categories (positive, negative or both) depending on the subjective sentiment associated with them. For this analysis, one particular adjective, *bohatý* 'rich', has been excluded, as this could otherwise automatically drive differences between the income groups (since the antonym *chudý* 'poor' is not included in the lexicon). The income groups are taken from the World Bank classification of economies of June 2019 ([4], [19]). The countries are categorized into four groups: High, Upper-middle, Lower-middle or Low Income countries. Further, only those nationalities that have been registered as staying for more than 90 days in the Czech Republic since 1994 [20], and the Czechs themselves, were included in the dataset for this study. Such an official list – of nationalities that have had a long-term presence in the country – points to the nationalities that may be present in any news text corpus from the country in focus.

The subcorpus used contains a majority of nationwide daily newspapers, regional editions from Bohemia and Moravia, and several news- and lifestyle-related magazines. The total number of titles is almost 200 from the very end of 1989 to the end of 2018, although there is reasonable coverage mainly from 1991 onwards. For more details and specific titles, see [18]. Because the SYN-series corpora are updated on a yearly basis, this method is repeatable.

[2] The material for this study is accessible at: https://www.su.se/english/profiles/irel5167-1.364672.

For this article, the material is a subset of 4,408,853 data points (observations) extracted from the subcorpus described above. It includes all the countries and nationalities that co-occur with the adjectives also described above. Certain variables were chosen to accompany the linguistic variables Noun and Adjective. Table 1 shows examples of variables and values from the Low Income group, containing co-occurrences of *vážný* + *Severní Korea* ('serious/significant + North Korea'), *důležitý* + *Afghánistán* ('important + Afghanistan'), *milý* + *Ugaňan* ('nice + Ugandan man') and *zabitý* + *Syřanka* ('killed + Syrian woman').

sent	**neg-adj**	**noun**	**adj**	**fq**	**adjfq**	**title**	**pub-year**	**form**	**wb-income**
POS/ NEG	A	severní korea	vážný	1	0.25	Hospodářské noviny	2006	COU	WB-LI
POS/ NEG	A	afghánistán	důležitý	1	1	Týden	2012	COU	WB-LI
POS	A	uganďan	milý	1	1	Deníky	2007	MASK	WB-LI
NEG	A	syřanka	zabitý	1	1	Právo	2016	FEM	WB-LI

Tab. 1. Example of variables and values from the dataset for the Low Income group

The explanatory variables [23, p. 7] include the following:

- **sent** for whether they have positive or negative sentiment – or both,
- **negadj** for whether the adjectives in the corpus source text are negated (N) or not (A),
- **fq** for the number of occurrences in one and the same issue of a newspaper or magazine,
- **adjfq** for the distance-adjusted frequency of each adjective-noun co-occurrence, see details under Method,
- **title** for the name of the newspaper or magazine where the co-occurrence is found,
- **pubyear** for the year of publication,
- **form** for country noun (COU) or nationality in masculine (MASK) or feminine (FEM) form,
- **wbincome** for World Bank income classification: High (HI), Upper-middle (UMI), Lower-middle (LMI) or Low (LI) Income.

The adjusted frequency is explained in more detail under Method below. The reason for adding a variable for the adjectival negation (A for non-negated and N for negated), is that one type of negation in the Czech language is expressed by a prefix: *moderní* 'modern, trendy' is negated as *nemoderní* 'old-fashioned, outdated'. The meaning changes [21], and the sentiment is interpreted here as the opposite. In the corpus, the lemma is the same for both word forms, which is why a distinction had to be made.

3 METHOD

3.1 Calculating Sentiment Values by category

Before creating Figure 1 below, a variable called Sentiment Value, "sent-value", was added, created by

- representing co-occurrences with a positive sentiment by the number 1, a negative sentiment by −1, and those classified as both by 0,
- multiplying this by the adjusted frequency, and then
- calculating the mean of the resulting number per year per group.

The adjusted frequency is calculated as follows:

$$\text{adjFq} = \text{SUM}\ (\text{fq} \times 1/\text{distance})$$

This is a sum of the proximity (following definition 2.11 in [22, p. 361]) of the adjective and the noun, weighted by the inverse value of their distance. Hence, co-occurrences that are close get a higher number, and those farther apart get a number that is low enough to represent their assumed (non-)prominence in the original news text. This is done to measure the potential influence of the adjective on the node noun, based on findings about proximity in Czech [22] (in particular chapters 2 and 5). In Table 1 above, the co-occurrence of *vážný* 'serious/significant' with *Severní Korea* 'North Korea' has an adjusted frequency of 0.25, which means that the noun (or bigram) and adjective are 4 words apart, and are thus given lower relevance when calculating the assumed sentiment for these nouns. The co-occurrences of *důležitý* 'important' with *Afghánistán* 'Afghanistan' and *milý* 'nice' with *Ugandan* 'Ugandan man' have an adjusted frequency of 1, which means they are adjacent to each other and thus assumed modifiers. In short: the lower the adjusted sentiment frequency, the lower the impact on the results.

Low values are included since they may nevertheless form part of the general discourse concerning the sentiment expressed in texts about the specific nouns, but they are given less weight. The adjusted frequency is utilized, rather than searching for pre- or post-modifying adjectives, since there may well be more than one word between the adjective and its modified noun in Czech, and adjectives in the near vicinity may also be highly relevant for this analysis. The Czech SYN-series corpora are not (yet) syntactically annotated, which means that this is the best approximation available.

In order to obtain a normalized number, the mean is finally calculated per year. This is needed since the number of data points per year differs (see "Composition of the corpus SYN version 8" graph in [18]). The smaller the sample, the more clearly visible the changes in sentiment value. A few outliers [23, p. 9], sometimes even

a single one, may change the total in a more visible way when the number of data points is low. For comparatively small samples, such as the Low Income group, which only makes up 2.5 percent of the whole dataset for this study, singular events may thus cause extreme peaks and troughs. Smoothed lines are, therefore, chosen for the graphs to make it easier for the human eye to interpret the differences over a long time span.

4 ANALYSES

World bank category	*Number of data points*
High Income	3,001,847
Upper-middle Income	990,576
Lower-middle Income	306,052
Low Income	110,378

Tab. 2. Number of data points in the dataset that are classified into each income group

Countries and people from countries that are classified as having high income represent such a large proportion of the dataset, 68 percent, that they form an ingroup by themselves, which is seen in Table 2. It is also apparent that the number of data points follows other income levels in the set.

4.1 Adjectives over the whole period

Aggregated over the whole period, there are certain patterns in the type of adjectives that are attributed to different income groups. This is shown in Table 3, where the most frequent adjectives are shown per income group.

High Income	*%*	*Upper-middle Income*	*%*	*Lower-middle Income*	*%*	*Low Income*	*%*
velký (‘large, big’)	35.82	velký	34.76	velký	36.57	velký	29.08
dobrý (‘good’)	17.85	poslední	16.86	poslední	17.08	poslední	16.58
poslední (‘last, final’)	15.85	**dobrý**	15.23	**dobrý**	12.88	teroristický	9.26
silný (‘strong, forceful’)	5.59	silný	6.01	**mírový** (‘peaceful’)	6.13	**dobrý**	8.72
rád (‘happy, delighted’)	5.00	**možný**	5.37	silný	5.28	**mírový**	7.76
základní (‘fundamental, basic’)	4.30	důležitý	4.96	**možný**	5.01	válečný (‘war, martial’)	7.75

High Income	*%*	*Upper-middle Income*	*%*	*Lower-middle Income*	*%*	*Low Income*	*%*
možný (‘possible, feasible’)	4.20	špatný	4.39	důležitý	4.50	bezpečnostní (‘safety, security’ [adj.])	5.90
špatný (‘bad, wrong’)	3.88	**lidský** (‘human, humane’)	4.27	teroristický	4.26	**možný**	5.80
důležitý (‘important’)	3.85	**rád**	4.19	špatný	4.24	špatný	4.60
nízký (‘low, reduced’)	3.66	těžký (‘hard, severe’)	3.91	**rád**	4.05	**lidský**	4.56

Tab. 3. The ten most frequent adjectives 1990–2018 per income group. Percentage of these ten. **Bold** = positively classified, underlined = negatively classified, normal style = classified as both. Translations into English are shown the first time the adjective occurs in the table

A few words are needed about the most frequent adjectives to better understand the content of the corpus in general. The adjective that occurs most frequently with the highest income groups is *velký* ‘large’. This multifunctional adjective [24, p. 122] co-occurs with the highest income group 395,141 times in the dataset, which is slightly more than one third of the total of these ten, and 13 percent of the total number of adjectives for this group.[3] In the opposite corner of Table 3, the tenth most frequent adjective for the lowest income group, *lidský* ‘human’ or ‘humane’, co-occurs with that group 1,640 times, which is only about 4.5 percent of these ten, and 1.49 percent of the total for that income group. There is thus a difference in the number of adjectives between the different groups, which explains why the income groups must be studied separately.

Some highly frequent adjectives are present in all groups, but the less-frequent are of more interest here: *silný* ‘strong, forceful’ appears less frequently in the Low-middle Income group, and not at all in the Low Income group; *rád* ‘happy, delighted’ shows a similar trend; *mírový* ‘peaceful’ and *teroristický* show an opposite trend, only being present in the two lower income groups, and increasingly so. *Bezpečnostní* ‘safe, secure’ is only present in the lowest income group, which will be discussed under Conclusions.

We also notice that the three highest income groups all have two adjectives that can be both positive and negative, and that they also have four of each sentiment classification. The lowest income group actually has four negative and five wholly positive adjectives in the top ten, but the fact that the negatively classified *poslední*

[3] *Velký* and *poslední* ‘last, final’ (adjectives that are in the top 3 for all groups) are indeed also among the most frequent adjectives in Czech news language [25, p. 11 and 13].

'last, final' and *teroristický* have such a large share of the total makes them prominent in the discourse and creates a downward trend. What Table 3 shows is thus a prominence of words relating to war and peace for the lowest income group. This gives us a hint about the discourse surrounding the so-called war on terror that emerged at the beginning of the present millennium. It seems that there are some adjectives here that clearly have an "inscribed" attitudinal value, i.e., a stable negative value over time and context [26, p. 2], and that may create a clearly negative picture of the negative persona some nationalities are given in the Czech press during these years. Now let us scrutinize the trends for the four income groups.

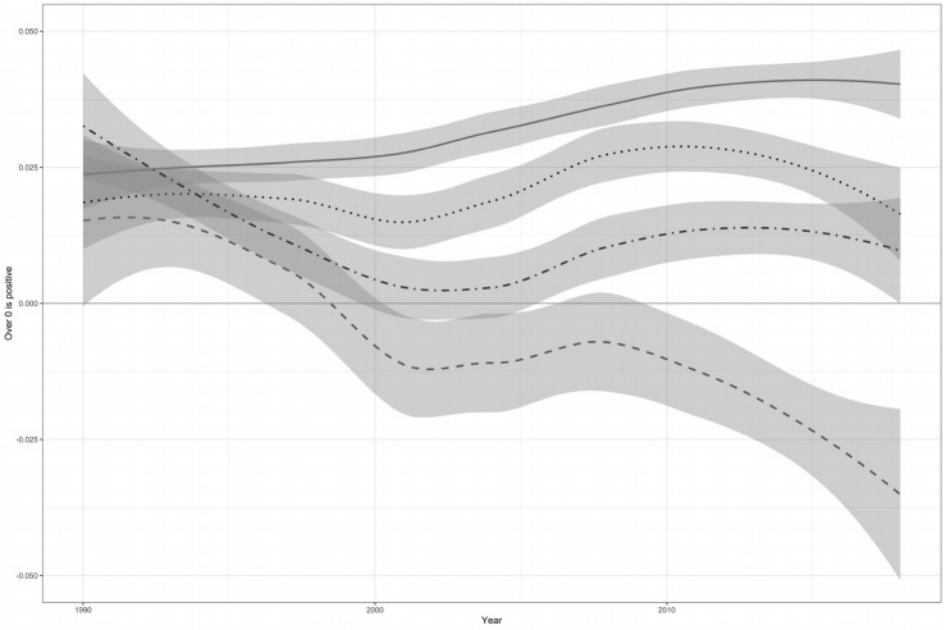


Fig. 1. Sentiments trends for the four income groups: High (solid —), Upper Middle (dotted …), Lower Middle (dot-dashed .-.-.) and Low (dashed - -) Income. The smooth, grey line equals the standard error with 95 % confidence interval. 0, which marks a balance between positive and negative adjectives, is shown on the horizontal line

The lower the income, the more negative the overall sentiment in Figure 1. As the corpus grows larger and more reliable in reflecting a general trend (from around 2004), this result becomes clearer, except possibly for the Upper-middle Income group, which closes in on the Lower-middle Income group towards the end of the period. The economic crisis in 2008 does not seem to have affected the three top income groups, whereas the trend for the lowest income group slumps from that year onwards. For this reason, the group with the lowest income is examined in closer detail.

4.2 Removal of an influencer

The Low Income group shows a clear trough around the year 2001. The time line recovers somewhat in the following years, but then deepens again around the year 2008. Let us, therefore, examine the noun + adjective co-occurrences that cause the dips during the period 2000–2009. This is shown by the adjusted senti-values in Table 4, visualized per year for the top and bottom ten of this group. It then becomes clear that one country and nationality alone creates the main reason for both the peaks and troughs in the Low Income group: Afghanistan. The nouns *Afghánistán 'Afghanistan'*, *Afghánec* 'Afghan man' and *Afghánka* 'Afghan woman' make up 34,115 of the 110,378 observations, i.e., 31 percent, for the entire period 1990–2018. It turns out that by far the biggest reason for the 2001 trough is the co-occurrence of the noun *Afghánistán* and the adjective *teroristický* 'terrorist [adj.]'. During 2001 and 2002, these two words in co-occurrence create an enormous trough, as seen by the sentiment numbers in Table 4. The co-occurrence returns in 2003, 2004, 2006 and 2007, although not as strongly.[4]

Most positive, Low income				*Most negative, Low Income*			
Year	*Noun*	*Adj*	*Total Adj Sent-v*	*Total Adj Sent-v*	*Noun*	*Adj*	*Year*
2001	afghánistán	mírový	34.4	-76.8	afghánistán	teroristický	2001
2001	afghánistán	možný	27.9	-41.0	afghánistán	poslední	2001
2002	afghánistán	mírový	24.7	-28.6	afghánistán	zvláštní	2001
2008	kongo	hluboký	20.1	-24.1	afghánistán	válečný	2001
2001	afghánistán	bezpečnostní	19.4	-19.8	afghánec	obyčejný	2001
2008	afghánistán	bezpečnostní	15.7	-13.8	afghánistán	tajný	2001
2002	afghánistán	dobrý	14.2	-12.8	afghánistán	těžký	2001
2002	afghánistán	bezpečnostní	14.1	-44.1	afghánistán	teroristický	2002
2003	afghánistán	mírový	13.9	-19.9	afghánistán	poslední	2002
2001	afghánistán	významný	12.4	-13.9	afghánistán	teroristický	2003

Tab. 4. The most positive and the most negative co-occurrences for the Low Income group 2001–2009

The negatively classified *adjectives are* countered by the positively classified *mírový* 'peace-, peaceful', and *bezpečnostní* 'secure, safe(ty-)', but the positive adjectives in Table 4 do not reach the same numbers. When added, as done above in Figure 1, the positivity of 34.4 (for *mírový*) plus 19.4 (for *bezpečnostní*) for the year

[4] This still holds even after removing the multifunctional adjectives [24, p. 122] for the negatively classified *poslední* 'last' and *obyčejný* 'ordinary, average', and for the positively classified *dobrý* 'good' and *možný* 'possible, believable'.

2001, and 24.7 (*mírový*) and 14.1 (*bezpečnostní*) for the year 2002 nevertheless does not make up for the negativity caused by *teroristický* in the same years.

This result might give reason to analyse the ingroup – Czechia and its people in a Czech context – as well. The outcome of that shows that in Figure 1, *Česko* 'Czechia', *Čech* 'Czech man' and *Češka* 'Czech woman' make up 547,332, i.e., 18 percent, of the total of 3,001,847 observations in the High Income group. The "obvious" ingroup of Czechs and their country is thus not as influential in their group as the outgroup noun of Afghanistan, and when removed, the High Income trend line nevertheless has a similar form.[5]

5 CONCLUSIONS

World Bank statistics rarely figure in linguistic articles, but their classification of income groups around the world has been used here alongside linguistic data to compare sentiment towards the countries and nationalities of these groups. Prominent parts of the overall rhetoric [1, p. 18] were indeed revealed when the data were scrutinized as a whole: the obvious ingroup of Czechs somewhat drove their income group's positive sentiment results, and an outgroup from a country far away from the ingroup drove their income group's negative results even more. Figure 1 thus shows that the higher the national income, the more positive the overall representation of the nationalities in these data. Even when the two most influential (most positive and most negative) noun groups are removed and the amount of source data swells around the year 2000, the sentiment trend lines remain separated.

This article argues further that when looking beyond the binary classification of positive and negative sentiment, a certain discourse emerges. As income decreases from High through the two Middle groups to the Low Income group, the adjectives *dobrý* 'good' and *silný* 'strong', 'forceful' decrease; *mírový* 'peaceful' and *teroristický* increase, and *bezpečnostní* 'safety', 'security' [adj.] make an appearance. Despite several of these being classified as positive in the Subjectivity Lexicon, they together reflect a prominent discourse about the lowest income group being involved in war and terrorism. This is a continuation of the "new security" discourse [5, p. 66] combined with the discourse surrounding the "war on terror" [1, p. 32] of the 2000s. This might even be a case of "canonical pairings" [12, p. 155]. The connection between peacefulness, safety and terrorism could form the basis of a future study on stereotypical conventions (such as "the poor are more likely to face negative than supportive behavior", [27, p. 243]). Such study could also compare the findings of Cvrček and Fidler [28, p. 17] where the noun "migrant" is closely associated with certain regional identities and the adjective "violent".

[5] This figure can be found at: https://www.su.se/english/profiles/irel5167-1.364672 together with other data from this study.

That an obvious ingroup, in this case Czechs in Czech newspapers, forms a large part of the peak (more positive sentiment) in such a combined volume of text could be expected. Without them, the High Income trend nevertheless remains clearly positive. That a single co-occurrence (of the noun Afghanistan and the adjective terrorist) creates such a large part of the trough in the trend line for the Low Income group constitutes a reason to examine the details, but also to plan future studies to explore how certain low-income nationalities have been treated in Czech media, or perhaps even Slavic media in general. It seems to be a case of resonance (as in [29]), i.e., repeating of linguistic formulae whether or not the next author agrees with the connotation given to the formula(e) by the previous author. Another next step could be to dig deeper into the co-occurrences of some of the nouns or adjectives, without the restriction of the Subjectivity Lexicon. This resonance could be explored even further with current-day corpora of both written and spoken text.

ACKNOWLEDGEMENTS

The author would in particular like to thank Václav Cvrček at the department of the Czech National Corpus for his continuous assistance, including the extraction of the dataset. Grateful thanks also go to the anonymous reviewers of this article, who kindly made clear what was clouded.

References

[1] Fairclough, N. (2006). Language and globalization. Routledge: Abingdon.

[2] Fowler, R. (1991). Language in the news: discourse and ideology in the Press. Routledge: London.

[3] Bednarek, M. (2019). The Language and News Values of 'Most Highly Shared' News. In F. Martin and T. Dwyer (eds.), Sharing News Online – Commendary Cultures and Social Media News Ecologies, pages 157–188, Palgrave Macmillan: Cham.

[4] World Bank Group. (2020). World Bank Country and Lending Groups. Datahelpdesk. Worldbank.Org.

[5] Wodak, R., de Cillia, R., Reisigl, M., and Liebhart, K. (2009). The Discursive Construction of National Identity. 2nd ed. Edinburgh University Press: Edinburgh.

[6] Lindqvist, A., Renström, E. A., and Gustafsson Sendén, M. (2019). Reducing a Male Bias in Language? Establishing the Efficiency of Three Different Gender-Fair Language Strategies. Sex Roles, 81(1), pages 109–117.

[7] Rönnbäck, K. (2014). The Idle and the Industrious – European Ideas about the African Work Ethic in Precolonial West Africa. History in Africa, 41, pages 117–145.

[8] Molek-Kozakowska, K., and Chovanec, J. (2017). Media representations of the "other" Europeans. Common themes and points of divergence. In J. Chovanec and K. Molek-Kozakowska (eds.), Representing the Other in European Media Discourses, pages 1–22, John Benjamins Publishing Company: Amsterdam/Philadelphia.

[9] Veselovská, K. (2013). Czech subjectivity lexicon: A lexical resource for Czech polarity classification. In K. Gajdošová A. and Žáková (eds.), Natural Language Processing, Corpus Linguistics, E-Learning, pages 279–284, RAM-Verlag: Bratislava/Lüdenscheid.

[10] Veselovská, K. (2017). Sentiment analysis in Czech. Ústav formální a aplikované lingvistiky (Institute of Formal and Applied Linguistics): Prague.
[11] Šindlerová, J., Veselovská, K., and Hajič, J. (2014). Tracing Sentiments: Syntactic and Semantic Features in a Subjectivity Lexicon. Proceedings of the XVI EURALEX International Congress: The User in Focus, pages 405–414.
[12] Paradis, C., Löhndorf, S., van de Weijer, J., and Willners, C. (2015). Semantic profiles of antonymic adjectives in discourse. Linguistics, 53(1), pages 153–191.
[13] UNICEF. (2020). Averting a lost COVID generation: A six-point plan to respond, recover and reimagine a post-pandemic world for every child. UNICEF Division of Communication: New York.
[14] Roberton, T., Carter, E. D., Chou, V. B., Stegmuller, A. R., Jackson, B. D., Tam, Y. et al. (2020). Early estimates of the indirect effects of the COVID-19 pandemic on maternal and child mortality in low-income and middle-income countries: a modelling study. The Lancet Global Health, 8(7), pages e901–e908.
[15] Tan-Torres Edejer, T., Hanssen, O., Mirelman, A., Verboom, P., Lolong, G., Watson, O. J. et al. (2020). Projected health-care resource needs for an effective response to COVID-19 in 73 low-income and middle-income countries: a modelling study. The Lancet Global Health, 8(11), pages e1372–e1379.
[16] Ziai, A. (2016). Development discourse and global history: from colonialism to the sustainable development goals. Routledge: London.
[17] Moretti, F., and Pestre, D. (2015). Bankspeak. New Left Review, (92), pages 75–99.
[18] Křen, M. (2019). Corpus SYN version 8 – Czech National Corpus Wiki.
[19] World Bank's Data Help Desk. (2020). How does the World Bank classify countries? World Bank Group.
[20] Czech statistical office. (2019). Cizinci v ČR podle státního občanství v letech 1994–2018 (Foreigners in the CR by citizenship in the years 1994–2018 (as of 31 December)).
[21] Kováříková, D., Chlumská, L., and Cvrček, V. (2012). What Belongs in a Dictionary? The Example of Negation in Czech. In R. Vatvedt Fjeld and J. M. Torjusen (eds.), Proceedings of the 15th EURALEX International Congress, pages 822–827, Department of Linguistics and Scandinavian Studies: Oslo.
[22] Cvrček, V. (2014). Kvantitativní analýza kontextu. Nakladatelství Lidové Noviny/Ústav českého národní korpusu: Praha.
[23] Brezina, V. (2018). Statistics in Corpus Linguistics: A Practical Guide. Cambridge University Press: Cambridge.
[24] Cvrček, V., Laubeová, Z., Lukeš, D., Poukarová, P., Řehořková, A., and Zasina, A. J. (2020). Registry v češtině (Registers in Czech). Nakladatelství Lidové Noviny/Ústav českého národní korpusu: Praha.
[25] Čermák, F., and Křen, M. (2011). A Frequency Dictionary of Czech. Routledge: Oxon.
[26] Don, A. (2016). "It is hard to mesh all this": Invoking attitude, persona and argument organisation. Functional Linguistics, (3), pages Article number 9.
[27] Lindqvist, A., Björklund, F., and Bäckström, M. (2017). The perception of the poor: Capturing stereotype content with different measures. Nordic Psychology, 69(4), pages 231–247.
[28] Cvrček, V., and Fidler, M. (2021). No Keyword is an Island: In search of covert associations (Preprint), pages 1–28.
[29] Du Bois, J. W. (2014). Towards a dialogic syntax. Cognitive Linguistics, 25(3), pages 359–410.

sciendo
DOI 10.2478/jazcas-2021-0061

KEY WORDS AND POLITICAL PARTIES IN THE 2020 PRE-ELECTION CAMPAIGN ON FACEBOOK

NATÁLIA KOLENČÍKOVÁ
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia

KOLENČÍKOVÁ, Natália: Key words and political parties in the 2020 pre-election campaign on Facebook. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 679 – 689.

Abstract: The research paper analyses key words found in pre-election communication of electorally successful political parties, based on which the main communication differences among those parties and the specifics of pre-election communication, as well as the pre-election discourse as a whole, are identified. Research material consists of political parties' microblogs published on individual political parties' Facebook profiles in the period from January 1, 2020 to February 28, 2020, with a reference corpus formed by the total of these microblogs. The analysis showed professionalization of political communication, the use of new, but also traditional ways of interaction with the electorate, pre-election communication based on the presentation of candidates, offensive and combative tone of the most successful parties, self-presentation, hints of persuasive and manipulative techniques, topic points of electoral programmes, but also thematic neutrality and non-specificity that suggest smaller electoral success.

Keywords: key words analysis, Facebook, microblog, pre-election communication

1 INTRODUCTION

This research paper focuses on key words in a pre-election campaign and is the part of a broader-based research, the aim of which is to map linguistic and communication specificities of the respective types of discourse. Although an interest in political and media communication has been noted in the Slovak linguistic environment (see more: [1], [2]), a more systematic focus on pre-election communication through social media is yet to emerge. The situation may be surprising, as social media offer relatively accessible research material, and with further analysis a more comprehensive picture of this specific communication sphere could be obtained. Today, this specific communication sphere falls under the field of political marketing [3], which also suggests further application possibilities for such research.

2 THEORETICAL AND METHODOLOGICAL BACKGROUND

2.1 Social media in pre-election campaigns

The importance of social media in the pre-election campaign is often mentioned in connection with the victories of U.S. presidents B. Obama (YouTube) and

D. Trump (Twitter). In the Slovak geopolitical space, the importance of such media is associated with the presidential election campaign of I. Radičová in 2009 (Facebook) and with the unexpected electoral success of the political party Sloboda a Solidarita ('Freedom and Solidarity') in 2010 (Facebook). The natural incorporation of social media into the pre-election campaign is determined by its specifics, as this type of media differs from the traditional electronic types such as radio and television.[1] Taking into account the topic of this research paper, what matters the most is that although anyone, even a political party, can be an author of any content, whether it is socially relevant or trivial and highly individualized, the basic technical and currently more lax legislative rules must be adhered to nonetheless. The communication model of disseminating the message through a network of participants strengthens the balance between the author and the recipient. This enables the voter to get an impression of active participation in political activities through being in virtual contact with a political party. The topics, which political parties covered in their microblogs or statuses in the pre-election period, therefore became a stimulating study subject. The aim is to identify thematic areas of individual political parties and, subsequently, to compare them while characterizing the main thematic areas of pre-election discourse. Topics of the specific political parties that can be perceived as prominent are the basis for the research.

2.2 Key words, target texts and reference corpus

A key words analysis (see more: [7], [8]) served as a functional and methodological support in meeting the objectives. A key word can be characterized as a unit "whose frequency within the text is significantly higher than it would be expected based on the frequency of this unit in the reference corpus" [9]. Key words based on using statistical tests (chi-square test or log-likelihood test) are corpus-based and they should represent general usage. The DIN (difference index) is calculated if units show a statistically significant difference. The range for this value is <–100, 100>. A value of –100 means that the given word form is not present in the target text, but can be found in the reference corpus, and a value of 100 means that the element is present in the target text, but is absent from the reference corpus.

The KWords application [10] was used for key words identification. However, this application does not identify lemmas, but word forms. Using a stop-list, pronouns, prepositions, conjunctions, and numbers were excluded from the corpora, and both statistical tests were selected, setting the minimum occurrence frequency to 5. The main focus was on key words with a DIN range of <80, 100>, as the main interest is on prominent political party topics, although other aspects were taken into account in the analysis as well.

[1] Specialised literature offers a number of approaches to social media and their specificities (see more: [4], [5], [6]).

The research material consists of communications obtained from the social network Facebook, posted on the profile pages of each electorally successful political party in the period before the Slovak parliamentary elections of 2020 (January 1, 2020 – February 28, 2020). This covered a total of 1,102 microblogs which consist of 66,517 tokens and 17,214 word forms, and together with the communication of the three selected political subjects (ranked by their electoral success: PS/SPOLU, KDH, SNS), they form the reference corpus (1,847 microblogs, 142,144 tokens and 23,293 word forms). The reference corpus was selected in accordance with the goal of this paper, which was to show the differences between the topics addressed by the individual political parties in the context of general social media pre-election standards applicable during the period under investigation. The next step could be to expand the material, i.e., to use larger (balanced) contemporary Slovak corpora to research social media communication. Table 1 shows the extent to which each political party participated in this number, as also the scope of the target texts is recorded for each political party. Microblogs were not analysed separately but as an integral part of text produced by each political party.

3 KEY WORDS ANALYSIS

3.1 Relevant data

Relevant data regarding the communication of individual political parties are presented before the actual key words analysis in Table 1. Outputs from the KWord application, together with statistical test values, as well as DIN values, are available on the GitHub platform due to space-saving efforts – https://github.com/NataliaKolencikova/Key-words-and-political-parties-in-the-2020-pre-election-campaign-on-Facebook. The examples used in other parts of the research paper are presented in an authentic form, without any corrections. Since the application was set to ignore case sensitivity, also proper names, if not quoted, are written with a lower-case initial letter.

Political party	Election result	Number of microblogs	Number of tokens	Number of word forms	Key words DIN = <80, 100>
OĽANO ('Ordinary People and Independent Personalities')	25.02 %	280	12,791	3,174	*kresťanská* 'christian', *nova* 'new majority', *kú* 'chu', *zámocká* 'castle', *zdola* 'from below', *sopko, bystriansky, moderuje* 'hosted by', *jožo, pročko, kristián, historickú* 'historical', *rozhodni* 'decide', *čekovský, jara, krajčí, rozprávať* 'talk', *peťo, naďa, spravodajstva* 'intelligence', *oborník, obyčajní* 'ordinary', *nezávislé* 'independent', *osobnosti* 'personalities', *posielajte* 'send', *debatovať* 'debating', *tieňový* 'shadow', *jaro, únia* 'union', *kandidátom* 'candidates', *súhlasíte* 'agree', *heger, dole* 'down'

Political party	Election result	Number of microblogs	Number of tokens	Number of word forms	Key words DIN = <80, 100>
Smer – SD ('Direction – Social Democracy')	18.29 %	119	5,064	1,808	*imigrantov* 'immigrants', *sd, zodpovedne* 'responsibly', *blanár, beseda* 'conference', *matovičovi, tlačová* 'press', *opozícia* 'opposition', *zodpovedná* 'responsible', *opatrenia* 'measures', *premier* 'prime minister'
Sme rodina – Boris Kollár ('We Are Family – Boris Kollár')	8.24 %	223	12,560	3,223	*holý, adriana, pčolinská, borisa, nachádzame* 'to be in', *miletička, krajniak, doplatkov* 'copayments', *lukáč, senica, pčolinský, boris, nájomných* 'rental', *poslankyňa* 'female parliamentarian', *rodina* 'family', *hnutie* 'movement', *bytov* 'apartments', *kollár, lieky* 'medicines', *odpovedať* 'answer', *ekonomický* 'economic', *vysielaní* 'broadcasting')
ĽSNS ('People's Party – Our Slovakia')	7.97 %	19	237	159	*kotleba*
Sloboda a solidarita ('Freedom and Solidarity')	6.22 %	252	13,455	3,933	*ideamedia, majerníková, priemyselná* 'industrial', *návod* 'guide', *zemanová, klus, oks, zelená* 'green', *ďuriš, nicholsonová, sas* 'fs', *sloboda* 'freedom', *solidarita* 'solidarity', *podnikať* 'run a business')
Za ľudí ('For people')	5.77 %	209	22,410	4,917	–

Tab. 1. Relevant data of key word analysis

3.2 OĽANO (Ordinary People and Independent Personalities, OPIP)

The results of the most successful political party, both the key words and their concordances, show that the entity frequently uses paid advertising for communication, about which the public must be informed if used in the pre-election campaign as it is a legislative obligation to do so. This is done through a short, usually formal addition to the microblog, which also has a given form: *Objednávateľ:* ***OBYČAJNÍ ĽUDIA*** *a* ***nezávislé osobnosti*** *(OĽANO),* ***NOVA, Kresťanská únia (KÚ), ZMENA ZDOLA, Zámocká*** *14, Bratislava IČO:42287511 Dodávateľ: Facebook Ireland Limited, 4 Grand Canal Square, Írsko, IČO:462962* 'Client: **Ordinary People** and **Independent Personalities (OPIP), New Majority, Christian Union (CHU), Change from Below**, 14 **Zámocká** street, Bratislava CRN: 42287511 Supplier: Facebook Ireland Limited, 4 Grand Canal Square, Ireland, CRN: 462962'. However, the relevant words that form this part of the microblog need to be taken relatively in connection to the parties' thematic pre-election campaign, as its competitors have no reason to use words such as the party's address, which justifies the statistical difference between these elements in target texts and the reference corpus. However, it can be noticed that the words

objednávateľ 'client' and *dodávateľ* 'supplier' are not key words nor words of statistically significant difference compared to the reference corpus of any political party, and, therefore, it can be stated that it is a common part of the examined discourse and that these administrative elements demonstrate the professionalization of political communication and are, in fact, common in pre-election microblogs.

The model and repetitiveness have an impact on some other key words that are related to wider technological possibilities of social media, such as live broadcasting, which serves as a tool to strengthen interaction with the voter – *CHCEME SA S VAMI* ***ROZPRÁVAŤ*** 'WE WOULD LIKE TO **TALK** TO YOU'*; Predstavujeme vám našich kandidátov do volieb 2020. Dnes sledujte naživo rozhovor s Erikom Ňarjašom, naším* ***kandidátom*** *č. 14. Svoje otázky na Erika* ***posielajte*** *do komentárov* 'We present you our candidates for the 2020 elections. Do not miss today's live interview with Erik Ňarjaš, our **candidate** no. 14. **Send** your questions for Erik to the comment section'*; Sledujte dnešnú diskusiu s (...)* ***Moderuje*** *Matúš* ***Bystriansky***. 'Watch today's discussion with (...) **hosted** by Matúš **Bystriansky**.'. These examples show that the aforementioned activities could be connected to a relatively large number of anthroponyms that refer to specific candidates (*sopko, pročko, čekovský, krajčí, naď, oborník, heger*). Taking into account that this political party was in the opposition in the pre-election period and those are politically rather unknown personalities. Their presentation by domestic subject has a significant impact on differences that occur when anthroponymic elements are compared to the general usage and what is more, colloquialisation discourse tendencies are indicated (*jožo, jaro/jara, peťo*). When specific candidates are presented, also specific terms appear among the key words. Even though their connection to political discourse is connotatively accurate, their full extent is only shown in specific collocations – ***tieňový*** *minister* 'the **shadow** minister' and *bývalý riaditeľ Vojenského* ***spravodajstva*** 'former director of Military **Intelligence**'. According to pre-election polls, this political party's leader name appears commonly in general usage, thus is absent from the group of strongest key words. Nevertheless, regarding the communication of his political party, the word *matovič* has a fairly high DIN (75). Moreover, it is collocatively associated with another key word – *Už dnes bude Igor Matovič* ***debatovať*** *s Richardom Rašim v relácii Na Hrane* 'Today, Igor Matovič will be **debating** Richard Raši on the Na Hrane show'.

Although the strongest key words do not include the word *voľby* 'elections', it is present in the OPIP communication by emphasizing the importance and uniqueness of Slovakia's situation, which should be the means of activating the electorate as it is based on criticism and attacks made against the currently ruling party; thanks to the elections, we will have ***historickú*** *možnosť zrušiť vládu mafiánov* 'a **historic** opportunity to abolish the mafia government' and *mafiu dáme vo voľbách spoločne* ***dole*** 'we will put the mafia **down** together in the elections'. *Rozhodni* 'decide' and *súhlasíte 'agree'* are the key words of the unique survey, in which people had the opportunity to vote on the points of the next government of the Slovak Republic mission statement.

3.3 Smer – SD (Direction – SD)

In the political communication of Direction – SD, the key word *imigrantov* 'immigrants' reaches the highest DIN and based on its occurrence in mainly negative contexts, it is safe to imply the use of emotionally based persuasive techniques. This creates a sense of threat and reinforces the fear coming from an idea that political elites may shift the attention from ordinary citizens to immigrants – *Zodpovedný plán lídra strany SMER – SD Petra Pellegriniho je o spokojnosti našich občanov, nie imigrantov* 'The responsible plan of the Direction – SD leader Peter Pellegrini is to secure well-being for our citizens, not immigrants'. The solution to this situation is responsibility, both on the voters' side (*Voľte* ***zodpovedne!*** 'Vote **responsibly**!'), as well as on the side of the state officials (*Migračná kríza je najväčšia výzva pre EÚ a preto je dôležité, aby sa aj budúca vláda správala* ***zodpovedne*** *a chránila naše hranice.* 'The migration crisis is the biggest challenge for the EU and it is, therefore, important that the future government behaves **responsibly** and protects our borders.'). In contrast to the disunited and unstable opposition (***opozícia*** *je rozhádaná* 'the **opposition** quarrels among themselves'*;* ***opozícia*** *vytvára chaos* 'the **opposition** creates chaos'*;* ***opozícia*** *bojuje medzi sebou* 'the **opposition** is full of infighting'), the political party itself should represent the desired responsibility (*zodpovedná zmena* 'responsible change' as an electoral slogan).

The key word *sd*, i.e., an acronym for social democracy, expresses the party's political orientation and is semantically related to the key word *(sociálne)* ***opatrenia***, (social) **measures**. This key word indicates that the abbreviated form of the name is common within general usage and also that this political party perceives social and media communication as public and formal. The dominant genre of the pre-election campaign is ***tlačová*** **beseda** '**press conference**', while the usage of extended technological possibilities of social media is outside their main focus. This is surely related to the main focus being on the electorate, whose main information source is still traditional electronic media.

The formal nature of communication is also supported by the fact that unlike political communication of OPIP, informal forms of politicians' names were not found in the Direction – SD communication. As this political party was the ruling party in the pre-election period, criticism from other entities is expected and, therefore, elements referring to its main representatives were not found among the strongest key words (*fico*, without statistically relevant difference; *pellegrini*, DIN = 76). As it is an already well-known political party, elements whose statistical significance would refer to activities aimed at the presentation of its members are excluded from the subject's communication. The only proper-name element that contradicts this statement is the key word *blanár*. However, the political party presents its leaders through their functions – *premier (Peter Pellegrini)* 'Prime Minister (Peter Pellegrini)' (*predseda* 'chairman', DIN = 72). Emotional and expressive expressions are used in the communication of then ruling party in order

to express a markedly offensive, critical and hostile attitude towards the leader of a competing political party – *Matovičovi sa nedá veriť ako premiérovi. Ohavná pravda o Matovičovi!* 'Matovič cannot be trusted as a prime minister. The disgusting truth about Matovič!'. The overlapping of emotionality with the aforementioned formality does not necessarily suggest erroneous interpretation, as by using key words analysis, it was possible to point out the already known hybrid nature of Internet texts that is caused by the disruption of the boundaries between stylistic factors that are otherwise clearly defined and objective [11].

3.4 Sme Rodina (We Are a Family)

The derived key words from the political party We Are a Family suggests they are using social media along three basic lines. The first is the line of candidates presentation (*holý, adriana, pčolinská, krajniak, lukáč, pčolinský)*, including their leader (*boris/borisa, kollár)*. As in the case of previous political parties, here, too, the emphasis on the candidate position was observed but this time within the party itself – *náš **ekonomický** expert Štefan Holý* 'our **economic** expert Štefan Holý'. The We Are a Family is the only political party that has a feminine word form that refers to a candidate among its key words – *poslankyňa* 'female parliamentarian', which points to a higher number of women in the leading positions, however the issue of gender-balanced language usage deserves much deeper attention.

The second thematic line is related to the use of social media opportunities as a means to communicate with the electorate and is based on a strong interactive "meet and greet" campaign. When the political party representatives met with the citizens during the campaign, they informed about it promptly and flexibly via social networks – *Momentálne sa **nachádzame** v meste Košice, a tešíme sa na každé jedno stretnutie, podanie ruky, či rozhovor* '**We are** currently **in** the city of Košice, and we are excited to meet, handshake and interview any of you'. The *miletička* market and the *senica town* are specific places that this political party visited more often than their political competitors. However, We Are a Family also draws on the new technological possibilities of social media and the inherent two-way communication chain on which it is based – *Boris bude **odpovedať** na vaše otázky už dnes o 20:00 v živom **vysielaní** tu, na našom Facebooku!* 'Boris will **answer** your questions today at 8PM in a live **broadcast** here, on our Facebook page!'.

The third thematic line is linked to the specific points of the election programme as it concerns topics of ***doplatkov za lieky*** '**copayments** for **medicines**' and ***nájomných bytov*** '**rental apartments**', which make evident populist orientation, but populism in this context is not understood as a persuasive or manipulative technique, but as a political approach based on connecting to the average person who may feel overlooked by the political elites [12]. The key word *rodina* 'family', which could intuitively also be one of the main points of the election programme, is contextually linked to the party presentation itself (*Ak by sa voľby konali v prvej*

*polovici januára, tak by naše **hnutie** SME **RODINA** podľa agentúry Focus volilo 7.1% voličov* 'If the elections took place in the first half of January, our **movement** WE ARE A **FAMILY** would receive 7.1% of the votes according to FOCUS agency') and to the administrative nature concluded in the OPIP key words analysis.

3.5 ĽSNS (People's Party – Our Slovakia, PPOS)

In the case of the political party of PPOS, only one anthroponym key word referring to its leader was identified. This may be due to lower number of communications published in the pre-election period as in 2017, Facebook suspended the main page of this political party with more than 80,000 followers for spreading inappropriate content. Although the party has re-created its official website, its communication with supporters is currently taking place mainly through dozens of fan pages and sites of regional organizations. Unfortunately, it is therefore not possible to draw more general conclusions about this party's thematic orientation. It is an unfortunate situation as these findings could be very interesting given the party's specific position on the Slovak political spectrum. Since the party repeatedly publishes mainly quotations of its leader, it can only be stated that his personality represents all party's opinions, as well as its current and future attitudes – *Marian **Kotleba**: Vybuchnutý panelák v Prešove ukazuje neschopnosť politikov KDH* 'Marian **Kotleba**: An exploded apartment building in Prešov shows the incompetence of the Christian Democratic Movement'.

3.6 Sloboda a Solidarita, SaS (Freedom and Solidarity, FS)

Paid advertisement is also frequently used by the FS political party as five of the fourteen key words falls under the aforementioned administrative side of microblogs – *ideamedia, majerníková, priemyselná* 'industrial (street)', *sloboda* 'freedom', *solidarita* 'solidarity'. In self-presentation, this party mostly uses the FS acronym, and the key words show a significant inclination towards Občianska konzervatívna strana ('Civic Conservative Party') (*PODCAST SaS: Kandidáti **OKS** sa ocitli na kandidátke **SaS** už v roku 2016, neskôr aj v eurovoľbách.* 'PODCAST FS: **OKS** candidates found themselves on the **FS** list of political candidates in 2016 and later in the European election.'). The FS party uses social networks to present its candidates to a limited extent (*klus, ďuriš, nicholsonová*) and the name of its leader is absent among the strongest key words. Although the key word *sulík* is thematised in the general usage, its DIN is 73. Using all technological possibilities of social media leads to one element appearing among the key words that refers exclusively to one candidate, Anna Zemaníková, who posts regularly on her profile page *Anna **Zemaníková** – **Zelená** pre naše deti* 'Anna **Zemaníková** – **Green** for our children', in which the party also claims to have environmental experts among its members. The FS is also characterized by drawing attention to their precisely described election programme, from which the economic area especially comes to the fore as it

is also included in the party's election slogan – *Náš **Návod** na lepšie Slovensko. 1144 konkrétnych riešení aby sa tu oplatilo pracovať, **podnikať** a žiť* 'Our **guide** to a better Slovakia. 1,144 specific solutions for making Slovakia a great place for work, **business**, and living'.

3.7 Za ľudí (For People)

There were no key words detected in the communication of the political party For People under the set criteria. However, this is not a matter of limited quantity of research material as in the case of the PPOS political party. This political party in its election campaign simply does not thematise any problem to an extent where it would significantly differ from the general usage. However, claiming that the analysis did not reveal any key words would be incorrect, as the strongest key words have a DIN of 72. Although a few of these key words suggest that the party's prioritised pre-election strategy was presenting its candidates (such as *remišová, valášek, šeliga*) or using new electronic genres (such as *spotify, audiomapa* 'audiomap', *epizóde* 'episode'), nothing provides more information about its programme orientation. Based on these findings, it can be assumed that this thematic neutrality contributed to the low election result of the party. However, confirming this hypothesis would require further key words analysis of political parties that did not fare well in the parliamentary elections, as it is not examined here due the limited length of this research paper.

4 CONCLUSION

Based on the analysis of key words obtained from the parliamentary parties pre-election communication on Facebook, general characteristics of Slovak pre-election communication and pre-election discourse can be drawn. The key words that were obtained through the aforementioned coherent and logical schemes confirm frequent use of paid advertisements, which suggest professionalization of political communication [4] that is, paradoxically, closely linked to administrative elements that can be found at the superficial and deep level of microblogging. The thematisation of political parties' specific areas in comparison with the general usage may be largely influenced by the model and repetition that can be observed in electoral slogans or in the use of fixed text schemes that are only changed in specific situations. This concerns mainly texts that are usually associated with the technological possibilities of social media and how they can be used, which is characteristic especially of political parties oriented towards a younger electorate (especially OPIS, We Are a Family, sometimes FS). However, the use of traditional election campaign strategies was detected as well (thematisation of press conferences and "meet and greet" campaigns). Political parties take into account the principles of the Slovak electoral system and accordingly present their candidates on social

networks while emphasizing their specific function or position. The form of their names, especially in the case of the most successful political party, indicates colloquialization tendencies. As for the leaders of political parties, their names can be found in the group of strong key words, but not the strongest, which indicates their partial thematisation in the general usage. However, this statement does not apply to the main representatives of We Are a Family and PPOS. Regarding the two electorally most successful political parties (OPIP, Direction – SD), an aggressive and combative tone of communication was detected whereas self-presentation is characteristic for the political parties We Are a Family and FS. The election programme main topics and strategies can be seen in the key words as well (survey – OPIP; immigrants – Direction – SD; copayments for medicines and rental apartments – We Are a Family; business – FS), however, according to this research, thematic neutrality and non-specificity implies lower electoral success. The thematisation of what might be considered a higher, noble, and more abstract idea can only be observed within the two most successful political parties' communication (elections as a historical chance – OPIP; responsibility – Direction – SD).

The key word analysis undertaken provided a more detailed insight into the complex picture of this unique communication sphere and helped to point out the differences related to thematic orientation of electorally successful political parties' pre-election activities, which may suggest its application possibilities within the field of political marketing. The discourse analysis outlined other possibilities for further research. What seems to be valuable and helpful is comparing results from this research with the results obtained from the key words analysis of electorally unsuccessful political parties, but also comparing key words with thematic words. A future focus on persuasive and manipulative strategies, especially on emotionally based techniques, *ad personam* attacks or populism, could also be pragmatically interesting.

ACKNOWLEDGEMENTS

This research paper was created within the project supported by the Štefan Schwarz Fund and within the Science Grant Agency – project VEGA. no. 2/0016/21 *Dictionary of the Contemporary Slovak Language – 7th Stage (Compilation, unification and editing of the dictionary entries and related lexicological and lexicographic research)*.

References

[1] Rašová, D. (2013). Pragmatika jazykových javov v masmediálnej komunikácii. Kontrastívna štúdia na materiáli v slovenčine a v nemčine. Kraków: Spolok Slovákov v Poľku, 170 p.

[2] Štefančík, R., and Dulebová, I. (2017). Jazyk a politika. Jazyk politiky v konfliktnej štruktúre spoločnosti. Bratislava: Ekonóm, 193 p.

[3] Jaworowicz, P. (2016). Wideokomunikowanie polityczne w Internecie. Youtube i polskie partie polityczne w latach 2011–2014. Warszawa: Difin, 223 p.
[4] McQuil, D. (2007). Úvod do teorie masové komunikace. Praha: Portál, 447 p.
[5] Manovich, L. (2001). The Language of New Media. Cambridge/London: The MIT Press, 354 p.
[6] Page, R., Barton, D., Unger, J. W., and Zappavigna, M. (2014). Researching Language and Social Media: A Student Guide. London/New York: Routledge, 202 p.
[7] Adolphs, S. (2006). Introducing Electronic Text Analysis: A Practical Guide for Language and Literary Studies. London/New York: Routledge, 159 p.
[8] Scott, M., and Tribble, Ch. (2006). Textual Patterns. Key Words and Corpus Analysis in Language Education. Amsterdam/Philadelphia: John Benjamins, 203 p.
[9] Cvrček, V. (2017). Klíčové slovo. In CzechEncy – Nový encyklopedický slovník češtiny. Accessible at: https://www.czechency.org/slovnik/KLÍČOVÉ SLOVO.
[10] Cvrček, V., and Vondřička, P. (2013). KWords. Praha: ÚČNK. Accessible at: http://kwords.korpus.cz.
[11] Patráš, V. (2009). Sociolingvistické aspekty electronicky podmienenej komunikácie. Karviná: Slezská univerzita v Opavě, 297 p.
[12] Populism. (2021). In Oxford Dictionaries. Accesseible at: https://www.lexico.com/definition/populism.

sciendo
DOI 10.2478/jazcas-2021-0062

‘AND WE ARE STUCK IN ONE PLACE, MINISTER.’ A STUDY OF EVASIVENESS IN REPLIES TO FACE-THREATENING QUESTIONS IN SLOVAK POLITICAL INTERVIEWS ON SCANDALS (A COMBINED APPROACH)

JANA LOKAJOVÁ
Department of Languages and Social Sciences, Institute of Languages and Sports, Faculty of Mechanical Engineering, Slovak University of Technology, Bratislava, Slovakia

LOKAJOVÁ, Jana: ‘And we are stuck in one place, Minister.’ A study of evasiveness in replies to face-threatening questions in Slovak political interviews on scandals (A combined approach). Journal of Linguistics, 2021, Vol. 72, No 2, pp. 690 – 704.

Abstract: The phenomenon of political evasiveness in the genre of a political interview has been the focus of several discourse studies employing conversation analysis, critical discourse analysis and the social psychology approach. Most of the above-mentioned studies focus on a detailed qualitative analysis of political discourse identifying a wide range of communication strategies that permit politicians to ambiguate their agency and at the same time boost their positive face. Since these strategies may change over time and also be subject to a culture specific environment, the aim of this paper is to discover a) which evasive communicative strategies were employed by Slovak politicians in 2012–2016, b) which lexical substitutions were most frequently used by them to avoid negative connotations of face-threatening questions, and finally, c) which cognitive frames formed a frequent conceptual background of their evasive political argumentation. The paper will draw on a combination of quantitative and qualitative approach to the analysis of non-replies devised by Bull and Mayer (1993) and critical discourse analysis in the sample of five Slovak radio interviews aired on the Rádio Express. The selection of interviews was not random- in each interview the politician was asked highly conflictual questions about bribery, embezzlement or disputes in the coalition. Based on qualitative research of Russian-Slovak political discourse (2009) by Dulebová it is hypothesized that a) the evasive strategy of ‘attack’ on the opposition and ‘attack on the interviewer’ would occur in our sample with the highest prominence in the speech of the former Prime Minister Fico, and b) the politicians accused of direct involvement in scandals would be the most evasive ones.

Keywords: attack, CDA, corruption, evasiveness, face-threat, hedging, interview, media, metaphor, muzhik, scandals, social psychology, political discoure

1 INTRODUCTION

In social psychology, the ambuiguity of political replies to interviewers‘ questions has been extensively studied by Bull and Mayer who devised a classification of eleven non-replies based on 18 televised interviews with British politicians from the 1987

General Election. The main contribution of their research lies in their innovative approach to 'evasiveness' which allowed for quantitative description of an otherwise intuitive impression of indirectness in political communication and thus permitted a stylistic comparison of politicians' idiolects. In their analysis of evasive 'non-replies', the turns of politicians were divided into a selection of topical shifts which presented an overt or covert way of political manouevering from the information scope of a question. The socio-psychologic research of the UK political discourse of 1980s discovered that the most prominent strategy of political evading of the question was to 'make a political point' or 'attacking' the question in order to minimise the damage done to the positive face of politicians [1]. The questions of interviewers were viewed as jeopardizing only the positive image ('face') of the politician since it was assumed that by giving a consent to be interviewed, the politician already accepted a certain degree of infringement of his freedom of action [2]. The concept of face introduced into the analysis of social interaction by Goffman [3], later developed in pragmatic linguistics by Brown and Levinson [4] was also adopted by Bull and Elliott in a later research and a classification of face-threats in questions of a political interview was provided [5]. In the present paper, their typology of face-threatening questions (FTA) will not be drawn on; the concept of a face-threat will be looked at from the point of view of its interactional consequences, i.e. the types of evasive arguments employed by politicians in their response to questions (cf. 3.1), and from a linguistic point of view (cf. 3.3), so a number of different strategies of an FTA lexeme replacement will be pointed out.

As in critical discourse analysis (CDA) evasiveness is viewed as one of misrepresentational strategies whereby the events of extralinguistic world are defocused through the usage of euphemisms, or implicit meanings [6], the political language in our paper will be examined also through the findings of cognitive semantic research of metaphorical language pioneered by Lakoff & Johnson [7]. It is assumed that a qualitative analysis of evasive replies of Slovak politicians might reveal analogies with the findings on the metaphorical language of the Russian political discourse, especially expressive forms of attacks on the media [8] or the employment of the image of 'muzhik' which was found as an effective means to model 'us' versus 'them' opposition [9].

2 METHOD

The five one-to-one broadcast radio interviews with Slovak politicians of the social-democratic government of 2012–2016 were chosen as a sample for the analysis of political evasiveness. The interviewees were questioned by the same interviewer, Braňo Závodský, who is known for giving his guests a hard time. The interviews revolved around many scandalous topics including a bribery case ('Bašternák') with an alleged involvement of the former Minister of Interior (R.K.), party nepotism in Regional State Bodies (R.F.), an unexpected coalition formation

involving both the Slovak National Party (A.D.) and the Hungarian party MOST-Híd, insider trading (B.B.), and the collapse of the bridge in Kurimany suggesting the wrong choice of contractor in the public tender (J.P.). The five interviews were transcribed from the youtube channel using Sonix software set up for the Slovak language. The timing of the interviews, the names of interviewees and the date of televised radio broadcastings are specified in Fig. 1[1]:

Interviewer / Politician	Braňo Závodský (Rádio Expres NAŽIVO)	
Róbert Kaliňák (R.K.)- Minister vnútra	Timing: 21.28 min.	17.11. 2016
Róbert Fico (R.F.)- Premiér	Timing: 23.21 min.	25.10. 2016
Andrej Danko (A.D.)- Líder strany SNS	Timing: 20.48 min.	5.9. 2016
Béla Bugár (B.B)- Líder strany MOST-Híd	Timing: 16.27 min.	20.6. 2016
Ján Počiatek (J.P.)- Minister dopravy	Timing: 12.01 min.	6.11. 2012

Fig. 1. Discourse event details

The computer aided transcription was subsequently manually cleared of any lexical or syntactic mistakes which were detected in turn-taking with rapid conversational exchanges and overlapping speech. As the goal of this paper was not to provide a detailed examination of evasive strategies in relation to their face threatening potential, the method chosen as the most suitable for our research was the combination of social-psychologic classification of non-replies with the CDA approach. Due to this, a chosen transcription method was not that of a detailed conversational analysis as adopted by Heritage or Jefferson [10].

In the first step, all 337 turns of the politicians were coded either as a) direct replies, b) intermediate replies or c) non- replies based on the method adopted by Bull and Mayer (1993) drawing on the definition of a question and reply by Quirk [11]. Direct replies were replies where the turn of a politician filled the missing information gap introduced by the question.

(1) IR: bude tu summit šéfov predsedníctva Únie

A.D.:. a to vzniklo znova z mojej iniciatívy a z udobrenia pána Schulza tak ako tu budú prvý krát premiéri budú tu prvý krát predsedovia parlamentov a Bratislavu dnes hľadá Európa na mape....

Intermediate replies were such answers of politicians which a) filled the missing information from the question in an indirect way, or b) provided incomplete information, or c) were interrupted and unfinished.

[1] https://www.youtube.com/watch?v=S9D_gab4rgE&t=1s, https://www.youtube.com/watch?v=XDYDF8_6rhk, https://www.youtube.com/watch?v=Eo8wXycHxUc, https://www.youtube.com/watch?v=tnT4tIpNjQg&t=271s.

(2) IR: dobre tá moja otázka, upriamoval som to na učiteľov, ale môžeme to kľudne rozšíriť aj na sestry, alebo aj na štátnych úradníkov...viac sa tam nedá?...lebo aj posledné sú signály také rozporuplné Minister financií hovorí nech si to rieši Minister školstva Minister školstva či zas má na to peniaze. ..je možné že od nového roka dostanú pridané alebo nie pán predseda?

R.F.: pán redaktor jediní učitelia majú v programovom vyhlásení vlády zakotvené zvyšovanie platov aj o konkrétne percentá

In the example above, the politician failed to provide a reply to a 'yes-no' question directed at nurses and employees in the public sector and instead continued to present a policy of the party related to a pay-rise of a different group of state employees, the teachers whom the interviewer excluded from the topical agenda of the question. Thus, the turn of a politician was tagged as an 'intermediate' and 'incomplete' reply, which was included as one of the eleven categories in Bull and Mayer methodology.

In 4.6% of cases, the distinction between a 'direct' and 'intermediate' reply of politicians was not clear-cut as the interviewer invited a politician to present his politicial view by a series of open-ended declarative questions. In those cases, an ethnomethodology had to be employed to resolve the fuzziness of the distinction, so that if in the upcoming turn the interviewer explicitly stated that the politician was evasive, twisted the events of the factual world, the previous turn of a politician was tagged as an 'intermediate' reply. In a 'non-reply' the policitian strayed from the agenda of the question to a full extent, e.g. by 'acknowledging the question', 'questioning the question', 'attacking the question' or 'declining to answer' (cf. Fig. 3).

(3) IR: ...a toto nie je prepojenie?
R.K.: no to je váš subjektívny názor [attacking the interviewer] na to vám už fakt nemám čo povedať [declines]

Thus, after all the intermediate and non-replies were identified in turns, they were segmented following the Bull and Mayer social-psychologic classification. However, as these categories were originally devised based on interviews from 1987, the original methodology was fine-tuned to include newly emerging strategies detected in our sample, e.g. a segment of 'expressing support' to colleagues, or the segment of 'referring to law'. All of these new strategies are discussed in section 3.2.

Finally, the five interviews were analysed using the CDA method focusing on a number of linguistic tools used by politicians to evade the questions asked, i.e. the

near-synonymical expressions substituting the face-threatening lexeme of the question (cf. 3.3) and the cognitive frames observed in the language of interviewed Slovak politicians when evading the questions (cf. 3.4).

3 RESEARCH

3.1 Frequency of evasive replies and their typology

Based on the methodology specified above, all 337 turns of politicians to questions asked were divided into direct, intermediate replies and non-replies and the percentage of evasiveness of each politician was calculated. The percentual distribution of evasive replies is specified below with the number of occurrences marked by (x).

POLITICIAN / TYPE OF REPLY	R.K. (SMER) Minister of Interior 2016-2018	R.F. (SMER) Prime Minister 2016-2020	A.D. (SNS) Speaker of Slovak Parliament 2016-2020	J.P. (SMER) Minister of Transport 2012-2016	B.B. (MOST-HÍD) Coalition partner 2016-2020
DIRECT REPLIES	37x	27x	13x	15x	21x
INTERMEDIATE REPLIES	46x	7x	7x	2x	16x
NON-REPLIES	54x	47x	18x	10x	17x
TOTAL NO OF RESPONSES (incl.non-replies)	137x (100%)	81x (100%)	38x (100%)	27x (100%)	54x (100%)
TOTAL NO OF EVASIVE REPLIES	100x	54x	25x	12x	33x
PERCENTAGE OF DIRECT REPLIES	27%	33,3%	34%	55.5%	38.8%
PERCENTAGE OF EVASIVENESS	**73%**	**66.7%**	**66%**	**44.5%**	**61.2%**

Fig. 5. Distribution of direct replies as opposed to evasive replies (intermediate replies and non-replies) to questions asked in five interviews

A quantitative analysis of intermediate replies and non-replies in our sample revealed that not all the politicians accused of scandalous affairs were equally evasive despite facing a possible vote of no confidence in the Parliament. The Minister of Interior (R.K.) accused of marring a criminal investigation in his own department was the most evasive one in his answers (73%) using an incomplete reply of 'starting but not finishing' indicating an adversarial style of an interview, along with the strategy of 'attacking' the opposition, 'attacking the interviewer', 'repeating the answer' and 'referring to law' which was a new strategy identified in the Slovak context (cf. 3.2).

The Leader of the Hungarian Party (B.B.) accused of making a fortune upon learning confidential information on the highway construction plans, was discovered to provide less evasive replies employing the same strategy of 'attack' and most importantly, a new strategy of 'referring' to religion', 'charity' or 'family' as possible excuses to accusations of FTA questions made by the interviewer (cf. 3.2)

The Minister of Transport (J.P.) proved to be the least evasive politician in our sample (44.5%) despite being accused of political responsibility for the deaths of workers in the construction project under a technical supervision of his Party nominees, using the strategy of 'declining to reply', or 'attacking the previous government'. A high number of replies provided in his interview resulted from the fact that from the middle of the interview the politician's argumentation started to be accepted by the interviewer as a clear, logical and satisfactory explanation of his decision not to give up on his office.

As illustrated in Fig. 3 below, in the superordinate category of 'making a political point' it was the category of 'giving reassurance' (8.75%), and 'attack on the opposition' (7.83%) which reached the highest frequency in the interviews. In the 2016 interview with a former Prime Minister Fico, the strategy of 'attacks on the opposition' was found with a lower frequency than expected; however, it was extensively used by Mr. Danko (10%) and Mr. Kaliňák (8.02%), both of whom used figurative language in delegitimation of their opponents, e.g. *obliať blatom* ('mud-sling'). In case of Mr. Bugár and Mr. Kaliňák, attacks on the opposition combined with those on the media suggesting the media collaborated with the opposition on discreditation of politicians or that the interviewer himself was involved in *mediálny lynč* ('media lynch'*)* of politicians.

(4)

R.K.: pretože toto má byť tá téma [impo not tackled] a pokračovanie toho mediálneho lynču…čiže [attacks interviewer]
IR: …ale no tak, zasa!

IR: ja viem ale tu sa na opozíciu vôbec nepýtam pán predseda
B.B.: ja viem [acknowledges] ale ja musím povedať kto vytvára takýto tlak? [returns Q] Kto hovorí občanom že fakticky všetci sú namočení do nejakých káuz? [attack]
IR: to občania si sami hovoria
B.B.: no no no. to tak nie je [Q based on a false premise] čítajú to čo niektorí tu podsúvajú [attack]

POLITICIAN / Evasive reply	R.K. (I.)	R.F. (II.)	A.D. (III.)	B.B. (IV.)	J.P. (V.)	PERCENTAGE (n=434)
2. Acknowledges the Q	4.94	3.36	1.66	1.42	-	3.22%
3. Questions the Q						5.06 %
a. Asks for clarification	1.24	2.52		1.42		
b. Reflects the question	3.70	2.52	3.33	7.14		
4. Attacking the Q						12.67%
Important Q not tackled	3.09	2.52	1.66	2.85	2x	
Q based on a false premise	4.94	5.04		5.71	1x	
Q hypothetical	1.24	3.36	6.6	5.71	2x	
Q objectionable	0.62	-		-		
Corrects the Q	2.47	-		1.42		
5. Attacks the IR	5.55	-	-	-		2.3%
6. Declines to answer	1.85	6.72	8.33	7.81	3x	5.52%
7. Makes a pol. point						45.62 %
Attacks	8.02	6.72	10	7.14	2x	(7.83%)
Presents a policy	-	6.72	6.6	2.85		
Gives reassurance	8.02	10.08	13.33	4.28	2x	(8.75%)
Appeal to nationalism	-	0.08	1.66	1.42		
Offers analysis	1.85	2.52		7.14	2x	
Self-justifies	11.73	5.88	18.33	2.87	2x	
Talks up	1.85	5.04	5	1.42	2x	
Justifies		4.20	1.66		1x	
i. Expresses support	2.46	5.88	5	1.42	-	
j. Refers to law	4.93	1.68	1.66	4.28	-	
k.Refers to religion	-	-	-	2.85		
l. Refers to charity	-	-	-	1.42		
m. Refers to family				1.42	2x	
8. Incomplete reply						13.82 %
Starts but does not finish	14.81	5.04	3.33	8.57		
Partial reply						
Half reply	1.24	1.68		2.85		
Negative reply	3.09	5.04	1.66	2.85		
e. Positive reply	0.62	0.08				
9. Repeats the answer	8.02	7.56	6.6	7.14	2x	7.6 %
10. States that Q has already been answered	3.7	3.36	3.33	4.28		3.45%
11. Apologizes		0.08		1.42		0.46%
12. Thanks		0.08				0.23%
TOTAL	100 %	100 %	100 %	100 %		100 %

Fig. 3. Distribution of evasive strategies (%) in five Slovak interviews adapted from Bull and Mayer (1993). The percentage of each strategy in each I–IV column reflects its frequency with respect to the total number of strategies in each interview; the percentage in the VI column reflects the percentage of each strategy with respect to the total number of strategies in all five interviews. In the first interview the total number of evasive strategies was n=162 (100%), in the second interview n=119 (100%), in the third one n=60 (100%), the fourth one n=70 (100%). Due to a low number of evasive replies in the fifth interview (n=23), the distribution of strategies in the V column was not stated in the percentage but in the number of occurrences (x)

The strategy of 'reassuring' the public was employed the most by Mr. Danko who as a new member of the government was given many questions on the future course of action and also by Mr. Fico who employed this strategy to boost the positive image of the party while at the same time attacking the negatively valued other- members criticising nepotism and calling for a change in 'autocratic' practices in the party.

(5) R.F.: no samozrejme tak potom je úplne normálne že môže dôjsť k takémuto prepojeniu budeme sa brániť ak bude niekto na Smer takýmto spôsobom útočiť bodka [reassures]

3.2 New evasive strategies

A close examination of five interviews with Slovak politicians revealed the occurrence of seven new evasive strategies (cf. Fig. 3). Five of them were subsumed under Bull and Mayer's 'making a political point' due to politicians' justification of political acts by reference to law (7j), religion (7k), charity (7l), family (7m) or by expressing a friendly support (7i) for the members states of the EU, or governmental colleagues. Other two strategies included a speech act of 'thanking' (12) and a 'positive reply' (8e). The highest number of references to law was noticed in the speech of Mr. Kaliňák and Mr. Bugár. Both of these politicians employed legalese language and resorted to quoting definitions based on the wording of the Constitution or the Commercial Code. In their answers, the argument about the autonomous ruling of judiciary, executive, and legal power was used when asked about the possible conflict of interest in their exercising of political function. Their resignation from a political office based on existing scandals was dismissed as premature or irrelevant.

(6) R.K.:..chráni...je to právo, ktoré máte z ústavy... [refers to law] to sú v podstate súkromné veci, ktoré robíte... čiže ja som neporušil nič....
.....
IR: s tým čo teraz kedy kedy to urobíte?
B.B.: to však samozrejme musí prejsť jednak koaličnou radou
IR: koaličnou radou
B.B.: a samozrejme musí byť aj aj v zákone to zabezpečené legislatívne [refers to law]

The benefit of legal argumentation was twofold- not only did it demonstrate the speakers' detailed knowledge of Slovak laws which were claimed to prove the business deals of politicians legal, but it also made the politicians appear as law-abiding citizens.

In the answers of Mr. Bugár, the strategies of 'referring to religion' and 'charity' were observed in the context of questions that implied that a politician took advantage of confidential information known to members of a government before a new highway project plan was oficially disclosed to the public. The politician was accused of a fraudulent way of obtaining money, an argument he dismissed by saying he was forced to sell the land near highway to the state and the only authority he should listen to is God.

(7)

IR: pretože sa o tom rozprávame
B.B.: pán redaktor viete čo v Biblii sa hovorí aby pravá ruka nevedela čo robí ľavá alebo naopak [religion] prepáčte prečo by som mal toto hovoriť [declines] je to moja vnútorná vec pred Bohom [religion] ja budem musieť zodpovedať a nie pred povedzme Matovičom

The money gained from selling the land was also indicated to be given to a non-profit organisation, such as the Church using the hedging of the conditional mood (*mohol by som povedať*) and the politician stated that he had formerly bought it from a family friend who was in need of money.

(8)

B.B.: nehovoriac o tom viete že na jednej strane sa zisťujú taketo blbosti [attack]
na druhej strane kľudne by som mohol povedať že polovicu som rozdal napríklad neziskovej organizácii [charity] napríklad cirkvi [religion]
B.B. …tomu aj ja rozumiem ale vtedy pred desiatimi rokmi od príbuzného ktorý sa dostal do ťažkej situácie potreboval peniaze [family]. Tak som kúpil pôdu.

In evasive replies of Mr. Fico, the 'support' of the governmental collegues accused of bribery and conflict of interest was stated and the members of Regional Administrative Bodies who expressed criticism of the party were attacked. The attack on the opposition merged with that on the media too, followed by a prefixed expressive verb *vykrikovať*.

(9) IR: ale to je také...
R.F.: no počkajte nie je to celkom pravda nie nie nie nie nie bolo to presne tak ako hovorím [reassures] preto treba byť absolútne opatrný pri vyšetrovaní akýchkoľvek káuz Robert Kaliňák má moju plnú dôveru je to jeden z najlepších ministrov vnútra [expresses support] aké kedy Slovensko malo [appeals to nationalism] ja nevidím dôvod teraz len pretože médiá opozícia vykrikuje konal [attacks]

A speech act of 'thanking' for the work done by governmental colleagues was also identified in the speech of Mr. Fico, who apart from using the strategy of 'negative reply' also employed a new strategy of a 'positive reply'.

(10) R.F. táto vládna koalícia je stabilná koalícia a chcem poďakovať koaličným partnerom za spoluprácu [thanks] to je moja odpoveď a nie je iná ako vy očakávate [repeats]

Unlike the 'negative reply' where the politician stated what he would not do instead of what he would do, in a 'positive' reply, the plan to undertake a certain course of action was described by a politician following a declarative question with the verb in negative form.

(11) IR: dobre vysvetľujem si to tak že tie zmeny sa nebudú týkať pána podpredsedu Kaliňáka
R.F.: nie, ja som povedal [implies Q asnwered] že budú zmeny na úrovni podpredsedov alebo v predsedníctve [positive reply] ale predsa len je to predovšetkým úloha na delegátov a nie pre nás dvoch teraz to riešiť [declines]

3.3 Linguistic substitution of FTA lexemes (CDA approach)

In our sample, a different representation of concepts was detected in hedged answers of politicians along with cognitive reframing suggesting coercion of the public and delegitimation of an opponent or media to boost a positive image of politicians. An FTA lexeme of the question taking the form of a) verbal and b) nominal phrases was replaced either by a) euphemisms to attenuate the impact of a face-threat by providing a more neutral equivalent, or on the contrary, b) substitution of a neutral lexeme from the question through figurative and dysphemic

language. As may be seen from the Fig. 4, the politician was depicted as a victim unjustly 'lynched'(R.K.) or 'slapped' (R.F.) by the media (*fackovanie*).

	FTA LEXEME	SUBSTITUTION	EUPH.	FIG.
NOMINAL PHRASES	kritika→	mediálny lynč (R.K.)		+
		fackovanie (R.F.)		+
	obvinenia→	vyhlásenia (R.K.)	+	
	prepojenie→	vzťah (R.K.)	+	
VERBAL PHRASES	odložiť→	vyriešiť (B.B)		
	vydržať→	byť pripravený na turbulencie (A.D.)		
	volať+ [IND. OBJ.]→	byť (+adverb) v Bruseli (R.K.)		
	vyšetriť→	prešetriť (R.K.)	+	
	zamlčať→	pomlčať (R.K.)	+	
	povedať→	potvrdiť (R.K.)		
	zostávať→	posúvať sa (R.K.)		
	odovzdať→	vziať (R.F.)		
	odvolať [IND. OBJ]	(ne) rezať hlavy (R.F.)		+
	ozvať sa [IND. OBJ.]+ adv [time] →	telefonovať (J.P.)		

Fig. 4. Substitution of FTA lexemes with noun and verbal phrases in the form of euphemisms (+) or dysphemic (figurative) language (+)

Although the figurative language was most frequently employed by the Leader of the Slovak National Party SNS (Fig. 5), it was also used by the former Prime Minister Fico in negative replies to evade the question on a possible removal of his Minister of Interior from the Government suggesting no guillotining of the ministers would take place (*nerezať hlavy*).

The highest number of verbal substitutions of FTA lexemes was made by Mr. Kaliňák who also produced the most evasive answers in our sample. Linguistic evasiveness of his non-replies lay in replacing the verbal prefix *vy-*, or *za-* by a different set of prefixes *pre-* and *po-*, which reduced the negative connotations of lexemes denoting his criminal inquiry (*vyšetrovanie)* or 'suppression' of the truth (*zamlčať*). The legalese meaning of 'investigation' was broadened and its neutral redefinition was amplified through a microprocedural narrative provided by the minister who included a story-telling description of individual speaking acts made by each 'accused' party during an alleged *vyšetrovanie.* A negatively valued meaning of an 'investigation' was thus substituted by verba dicendi (*povedať-porozprávať*).

Lexemes with positive connotations operating within the co-text of a politician's reply were also substituted; the interviewer uncovered and reframed only verbal predications of politicianswhich gave rise to unwanted implicatures depicting themas the ones who only postponed the solution of the problem (*odložiť*) instead of fixing it (*vyriešiť*), or implying unwillingness of the politician to cooperate as in *zostať na mieste* ('be stuck').

Although an implication that a co-partner in an interview misunderstood the provided argumentation presents a common rhetorical strategy of language manipulation [12] and evasiveness from the topic of the question might be hard to detect for interviewers [13], in the excerpt below praising of the interviewer, e.g. *posúvate sa o kúsok ďalej* ('you are gradually making progress') was dismissed by the journalist. The implicature that he was in fact slow in understanding the politician's argumentation was successfully decoded and challenged through a critical counter-argument of the interviewer claiming 'we are stuck in one place' (*zostávame na jednom mieste*).

(12) R.K. ja som nezamlčal…akože
tu sa presne posúvate opäť o kúsok ďalej…
IR: zostávame na jednom mieste
R.K.: ja som nikdy neklamal
IR: no ale ani ste nehovorili pravdu

3.4 Cognitive frames of evasive replies

In line with modern cognitive approaches to political discourse where a metaphor occupies the central position due to its persuasive character [14], the evasive replies of Slovak politicians were carefully analyzed looking for hidden conceptual frames enabling to reduce the political accountability through delegitimation or reference to certain practices as social norms (epistemic modality). Apart from generic clichés on politics as 'an art of a compromise', two main types of cognitive metaphors revolving around the concept of politics were identified in their speech, one of them anthropomorphic related to human activities (coffee-drinking) and the other one social (crime, fear, otherness) following the classification of Chudinov [15].

Politics as coffee-drinking and confidential business-making

In the discourse of the Slovak National Party Leader, politics was described as a serious deal happening over the cup of coffee, making it possible for certain politicians to have more personal discussion, which bound them to keep their word.

(13) A.D.: ale podstata celej tej nenahraditeľnosti spočíva v tom, že si nemáte s kým dať serióznu kávu…ako predseda politickej strany SAS si s vami dá kávu, povie vám niečo medzi štyrmi očami, vyjde pred médiá a bude klamať…

In contrast to Habermasian perception of the 17th and 18th century coffee-house as a new public sphere where critical discussions on state affairs were held among intellectuals [16], the metaphor of 'coffee-drinking' in the speech of the SNS party leader gained a private, secret-code meaning symbolizing the forging of a strong kinship. The confidentiality of political business deals was depicted as a standard practice through an implicature linked to the lexeme *predsa* ('in fact') used as an argument by the former Minister of Interior to evade the question on his not disclosing the information on a business made with Bašternák. The Minister stated 'in fact no Member of the Parliament speaks about his own things'.

(14)

R.K: že práve kúpou vzťah končí a nie začína...
to je základný rozdiel a keď som na to dostal
otázku, tak som na to jednoducho odpovedal..
proste vyložil som všetky karty na stôl...
R.K.: žiaden poslanec, ktorý príde do parlamentu
predsa nehovorí o svojich veciach…to je jeho vec.
tu je Igor Matovič...
IR: ale nemá problém s Bašternákom... na rozdiel
od vás..

Otherness of political opposition as homosexuality

The most figurative language expressions in our sample were discovered in the discourse of the leader of the Slovak National Party. In the 'us' versus 'them' ideological squaring [17] the political opponents were not only depicted in a visually emotive way using the metaphor of 'crime' inciting fear by an idiom *bodnúť do chrbta* ('backstabbing'), but also in a discriminatory way. They were tacitly attributed a homosexual identity by voice-qualifiers, e.g. *preskakuje hlas* ('trembling voice') or a 'woman-like' character (*zženštilosť*), which helped emphasise a stereotypical perception of queer men not acting as real men. The narrative of a strong alpha-male politician protecting the people from chaos thus points at similarities between the Russian national political discourse and Slovak national discourse where attribution of a homosexual label may be used as a form of delegitimation of the 'other' [18].

4 CONCLUSION

A combined social-psychologic and CDA analysis of the phenomenon of political evasiveness to questions in five selected one-to-one Slovak radio interviews revealed that the strategy of 'attacking' the opposition was a common technique of a topical shift in the Slovak context of 2012–2016. Although its frequency reached the second highest ranking (7.83%) of 'making a political point', it was not

associated most commonly with the political discourse of Mr. Fico as expected (6.72%), but with the Slovak National Party Leader who also used the most expressive language to delegitimise his political opponents (10%). Attacks on the interviewer were only made by Mr. Kaliňák. The hypothesis that politicians directly involved in scandalous accusations would be the most evasive ones was confirmed – the Minister of Interior accused of marring the investigation of his own case was the most evasive one of all interviewed politicians (73%) and the Minister of Transport accused of responsibility for deaths of workmen on a construction project delivered more satisfactory replies (44.5%). Apart from discovering new strategies of evading the questions, e.g. 'referring to law', 'religion', 'charity' or 'family', a cognitive frame of real 'manhood' was detected in the discourse of Slovak National Party Leader, A. Danko who implicitly depicted political opponents as homosexual, which indicates a similarity between Slovak and Russian national political discourse of recent years.

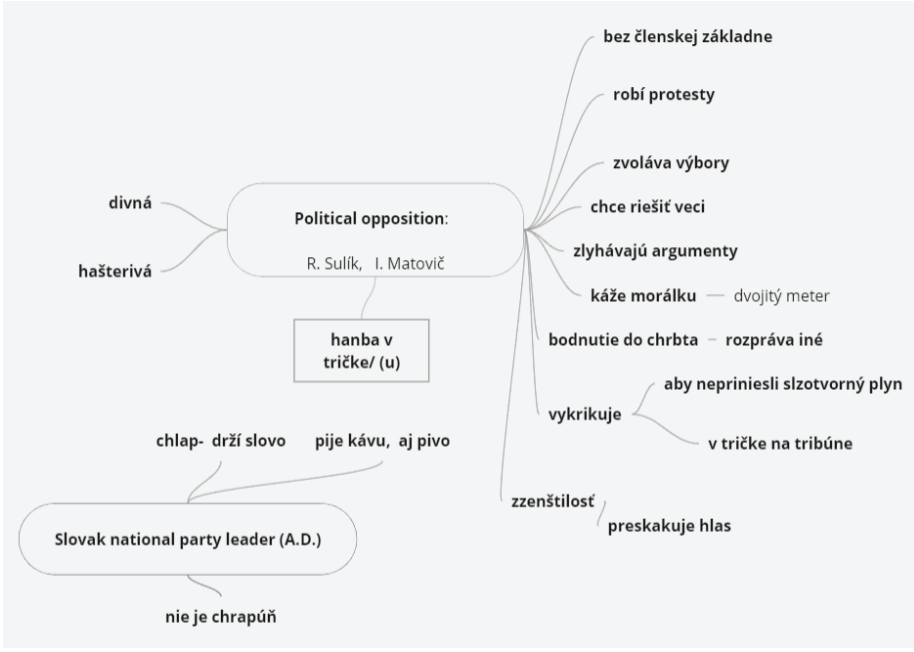


Fig. 5. Delegitimising the political opposition in the SNS Party Leader discourse

ACKNOWLEDGEMENTS

The research has been supported by the Cultural and Educational Grant Agency (KEGA) MŠVVaŠ SR under the grant 034STU-4/2019.

References

[1] Bull, P., and Mayer, K. (1993). How not to answer questions in political interviews. Political psychology, 14(4), pages 651–666. Accessible at: http:// www.jstor.org/stable/ 3791379.

[2] Jucker, J. (1986). News Interviews: A Pragmalinguistic Analysis. Amsterdam: Gieben.

[3] Goffman, E. (1967). Interaction ritual: Essays in face to face behaviour. Garden City, NY: Doubleday.

[4] Brown & Levinson in Huang, Y. (2013). Pragmatics. Oxford: Oxford University Press.

[5] Elliott, J., and Bull, P. (1996). A Question of Threat: Face Threats in Questions Posed During Televised Political Interviews. Journal of Community & Applied Social Psychology, 6, pages 49–72.

[6] Chilton, P. (2004). Analysing political discourse. Theory and practise. Routledge, 2004.

[7] Lakoff, G., and Johnson, M. (1980). Metaphors we live by. Chicago: The University of Chicago Press.

[8] Dulebová, I. (2009). Jazykové prostriedky manipulácie v modernom ruskom a slovenskom politickom diskurze. In Literatúra v kontexte slovanskej kultúry 20. storočia, pages 96–101, Banská Bystrica: UMB.

[9] Ryabova, T., and Ryabov, O. (2011). The real man of politics in Russia: On gender discourse as a resource for the authority, 42, pages 58–71.

[10] Have, P. (2007). Doing conversation analysis. A practical guide. London: Sage.

[11] Quirk, R., Greenbaum, S., Leech, G., and Svartivik, J. (1985). A Comprehensive Grammar of the English language. London: Longman.

[12] Edmüller, A., and Wilhelm, T. (2010). La Manipulation. L'art d'influencer à vôtre portée. Édition française. Ixelles: Les Miniguides Ecolibris.

[13] Gabrielsen, J., Jonch-Clausen, H., and Pontoppidan, Ch. (2017). Answering without answering: Shifting as an evasive strategy. Journalism, 21(9), pages 1355–1370.

[14] Charteris-Black, J. (2011). Politicians and rhetoric: the persuasive power of metaphor. London: Palgrave Macmillan.

[15] Chudinov, A. P. (2001). Rossia v metaforicheskom zerkale: kognitivnoe issledovanie politicheskoy metafory (1991–2000). Ekaterinburg: Ural.

[16] Habermas, J. (1989). The Structural Transformation of the Public Sphere. English translation. Cambridge: Polity Press.

[17] Dijk, V. (1998). Ideology. London: Sage.

[18] Gaufman, E. (2018). Money can't buy it? Everyday Geopolitics in Post-Soviet Russia. In Informal nationalism after communism. London: I. B. Tauris.

sciendo
DOI 10.2478/jazcas-2021-0063

LEXICAL BUNDLES IN THE CORPUS OF SLOVAK JUDICIAL DECISIONS

MIROSLAV ZUMRÍK
Slovak National Corpus, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia

ZUMRÍK, Miroslav: Lexical bundles in the corpus of Slovak judicial decisions. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 705 – 718.

Abstract: The paper follows the tradition of research in legal linguistics and into formulaic language, specifically into lexical bundles. The aim of the paper is to describe lexical bundles in samples from the corpus of Slovak judicial decisions OD-JUSTICE by means of quantitative characteristics of the identified bundles and by their comparison with bundles found in two other specialized corpora: the corpus of Slovak legal regulations and the corpus of annual reports by Slovak public institutions. For the identification of bundles, the concept of the h-point was used. Identified bundles are described with respect to their maximal, minimal, average, median and mode values, distributions and ratios. The aim of the paper is to outline an interpretation of these bundle characteristics with regard to communicative function(s) of compared document genres.

Keywords: lexical bundles, formulaic language, judicial decisions, specialized discourse, legal linguistics, pattern-driven research

1 INTRODUCTION

That a natural connection between language and law already exists can be deduced from the fact that language, according to J. Prusák, serves as a presupposition for the existence of the legal norms in which they are written [1, p. 295]. This relationship brings about another close connection, namely that which exists between law and linguistics. Their interaction, termed legal linguistics, can be seen as a "mutual arena for cooperation" [2], where one's interest in law may justifiably imply an interest in the nuances and peculiarities of legal language [3, p. 1]. Following the tradition of Slovak legal linguistics pioneered by Š. Luby or R. Kuchár, an interest in the linguistic aspects of Slovak legal language nowadays covers a range of areas, from the stylistic features of (mostly) legal norms ([3], [4]) to anaphoric and other tools for creating concise legal texts [5].

One phenomenological observation of legal language might be that legal texts tend to contain recurrent word sequences. This linguistic phenomenon has been studied as so called formulaic language, or formulaicity, which is often understood as characteristic of legal discourse [6, p. 7]. Such co-occurrence of language units

constituting multi-word expressions is, however, a feature present in the language in general – a systemic feature or an "axiom" [7, p. 10]. Formulaicity is further conceptualized as a pole on a continuous scale of qualities, as opposed to the pole of free creativity [8]. Given its scale-like nature, formulaicity can be studied and expressed quantitatively, similar to qualities of terms [9] or translated texts [10]. In other words, the formulaicity can be measured and therefore can be searched for possible "formulas" for formulaic structures [11], which can cover up to 40 types, from amalgams to unanalysed portions of speech [12, p. 3].

2 THE STUDY OF LEXICAL BUNDLES AND THE RESEARCH AIM

One of these formulaic structures in linguistics has been studied as so-called *lexical bundles* [13], such as '*on the other hand'*, '*as can be seen'* or '*it is recommended that'*. Researchers focus predominantly on more or less specialized discourses: academic discourse [14], medical leaflets [15] and legal texts, the latter from perspectives such as genre [16], linguistic structure [6], translation strategies [17] or legal semantics [18].

Lexical bundles have traditionally, by D. Biber, been defined on the basis of frequency as "sequences of word forms that commonly go together in natural discourse" and that "show statistical tendency to co-occur" [13, pp. 989–990]. Originally, three criteria have been used to define what classifies as a lexical bundle: these are, the minimal orthographic length of a sequence (3–8 tokens), its minimal normalized frequency (this cut-off point varies in individual approaches between 10, 20, 40 or more) and the "dispersion range" of a bundle throughout the individual texts in a given sample – this value could be, for example, the minimal appearance of a bundle in 5 different documents [13] or in 10 percent of sample texts [19].

Lexical bundles and other formulaic sequences are considered to be the building blocks of a given discourse ([20], [6]) which improve processing efficiency in communication [21]. One's command of such multi-word sequences is also said to have a sociological value for signalling an individual's belonging to a community [12] or a pedagogical value in teaching and mastering of specialized discourses [22].

Since the introduction of lexical bundles as a term, a methodological discussion has also emerged regarding a "fine tuning" of the process by which bundles are identified [23]. This involves searching for a method which would not only be based on arbitrary set cut-off points, but one that would uncover bundles more or less typical of a given discourse. Among such methods, Ł. Grabowski proposes the use of multiple sampling techniques, and, even where such bundles are identified as the most frequent and often structurally incomplete word sequences, he recommends that efforts to be directed at ways to identify preferably "structurally complete" bundles which can more easily be ascribed discourse function(s) [23, pp. 63–68].

Bearing the need for such methodological "fine-tuning" in mind, I would like to apply the method adopted in Cvrček et al. [24], where the formulaicity of texts is

expressed as the number of those 5-grams that have a frequency higher than the so-called h-point, introduced by I. I. Popescu [24, pp. 63 and 75]. The rationale for using the concept of h-point is that this point divides words into two areas: a branch of more frequent synsemantic words, and a branch of less frequent autosemantic words. If an autosemantic word, according to its frequency, appears in the synsemantic area, such a word can be perceived as "some kind of anomaly in comparison to 'neutral' texts" [25, pp. 217–218], that is, in some way specific to the text from which it originates. The h-point can be defined as the point where the rank of a given word equals its frequency. Where there is no such word, the h-point is calculated as follows, where $r(ank)_1 > r(ank)_2$ and $f(requency)_1 > f(requency)_2$:

$$(f_1r_2 - f_2r_1)/((r_2 - r_1) + f_1 - f_2))$$

By using the h-point calculations and quantitative characteristics, such as maximum, minimum, average, median or mode values, distribution of bundles throughout documents and corpora, as well as qualitative characteristics of the bundles (their communicative function), the aim of the paper is to describe lexical bundles in the corpus of Slovak decisions by means of comparison with bundles found in the corpora of Slovak legal acts and annual reports by Slovak public institutions. In other words, the paper is an attempt to find out whether, and to what degree, these quantitative characteristics correspond with the nature of both documents and institutional procedures from which the documents result.

3 DATA

The material basis for the research consisted of three specialized corpora:

1. The rather large corpus of Slovak judicial decisions **od-justice-1.0** (further referred to as **OD**).	**4,149,442,677** tokens	**2,622,795** documents
2. The corpus of Slovak legal regulations (**Korpus slovenských právnych predpisov k 2020-12-01, v1.4**, further referred to as **A(cts)**).	**43,750,050** tokens	**20,186** documents
3. The corpus of annual reports by Slovak public institutions **gov-vs-1.0** (further referred to as **AR**). Here, it should be noted that the data had been de-duplicated at the level of paragraphs, which needs to be taken into consideration.	**17,864,463** tokens	**1016** documents

Tab. 1.

Corpora 1 and 3 are available via the Department of Slovak National Corpus at https://bonito.korpus.sk, corpus 2 via the webpage of the Ľ. Štúr Institute of Linguistics at https://alica.juls.savba.sk).

Because the magnitude of corpora caused technical problems regarding the search for relevant bundles within its range, it was decided to work with two samples from each corpus, one smaller, at approx. 2 million tokens and one larger at approx. 10 million. The aim was to work with large portions of data of similar dimensions from all corpora.

In the corpus of judicial decisions, the verdicts can be grouped and searched in accordance with the date on which they were announced. This means that one can choose a subcorpus of decisions which only have this date in common. For analysis of decisions, I therefore picked 1000 documents from 10 July 2018 and 4968 documents from 29 June 2016.

Similarly, documents in the corpus of legal regulations are organized according to the year of their promulgation, but also according to subgenre (act, decree, resolution, regulation etc.). I decided to only work with acts, where the smaller sample were 472 documents from 2018 and 2019, and the larger sample of 2561 documents from the years 2007–2017 and 2020.

The annual reports are, understandably, sorted according to the years which correspond with the activities of the institutions they summarize. Here, the smaller sample consisted of 110 reports from 2012, while the larger one consists of 536 reports from 2007–2011, 2013 and 2014.

4 METHOD

From the 6 samples (two for each corpus, one smaller and referred to as OD/A/AR, one larger, referred to as ODext/Aext/ARext), it was necessary to identify lexical bundles above the respective h-point in each of the samples. Lexical bundles were identified using the CQL search in NoSketch Engine, as 5 consecutive tokens within a sentence (the latter condition so that the 5-token window would not take in sequences of words from neighbouring sentences). In order to further avoid counting punctuation, symbols or numbers as words, the tokens had to be attributed a morphological annotation in the range S (noun) – Y (conditional morpheme), thus excluding W (abbreviation, symbol) – 0 (number).

The identified 5-grams were ranked according to their absolute frequency. The next step was to find the h-point of a given sample: with ranks of individual bundles already identified, a search was made for the bundle that would have the same rank and the (normalized) frequency (*ipm*). In this way, the h-point for a given sample was either found directly (in 5 cases), or calculated using the formula mentioned in Section 2 (in 1 case).

The bundles above the h-point were sorted manually into groups, consisting of either

1. at least partially overlapping 5-word bundles, which constitute parts of longer bundles, as in *len do uplynutia lehoty na* 'only until expiration of (…) period'[1] – *do uplynutia lehoty na podanie* 'until expiration of (appellate period)', or

2. bundles with similar, albeit not identical wording, as in *Poznámka pod čiarou k odkazu* 'Footnote with respect to reference' – *Poznámky pod čiarou k odkazu* 'Footnotes with respect to reference'.

The issue of dealing with overlapping bundles is addressed in Grabowski [23, p. 63–67] where several possible approaches are mentioned. Here, the groups were identified manually by looking at the frequent right and left concordances of a bundle. The values of h-points and list of bundles above the h-point in each sample were then arranged in 6 tables for further processing, that is, counting and sorting overlapping and similar bundles within groups. Apart from bundles in groups (consisting of 2- to 12 bundles), those not contained within longer ones (thus representing genuine 5-grams) were also counted.

The 6 sample tables further contained values of normalized frequencies for each bundle, their dispersion throughout the documents, the ratio between the dispersion and number of sample documents in total. For these characteristics, the maximum, minimum, average, median and mode values were calculated. The values are then arranged in tables that serve as the starting point for the findings in Section 5.

5 FINDINGS

5.1 Maximal, minimal and average values of tokens per documents ratio in the samples (*tok/doc*)

	OD	**ODext**	**A**	**Aext**	**AR**	**ARext**
tokens	1,846,380	10,092,069	2,233,571	10,952,481	1,889,037	10,185,459
documents	1,000	4,968	472	2,561	110	536
max **tok/doc**	13,400	27,025	121,796	186,700	144,857	127,152
min tok/doc	239	249	109	97	1,932	1,652
avr tok/doc	1,846	2,031	4,732	4,276	17,173	19,003

Tab. 2.

The values vary in all three corpora, while the span is larger in the A sample. This makes the minimum values in A more similar to that of OD, while the maximum in A is more similar to AR. Even though the total numbers of tokens in both samples from all three corpora, as well as the ratio between token count in smaller and larger sample remain approximately the same (1:5), the number of documents in corpora differs, which makes average document length unequal. Comparison of smaller and larger samples shows that their average document length is approximately the same.

[1] I would like to thank Juraj Kotrusz for translation of some of the legal lexical bundles.

5.2 Number of lexical bundles above the h-point (*LBs > h-point*)

	OD	ODext	A	Aext	AR	ARext
LBs > h-point	106	85	55	42	33	26

Tab. 3.

This value remains approximately the same within each of the corpora (106 and 85 in OD, 55 and 42 in A, 33 and 26 in AR) while the ratio between those corpora can be described as 1 – ½ – 1/3. This would seem to indicate a higher degree of formulaicity found in the OD corpus, however, it needs to be considered that there are half as much documents in OD. The number of bundles in the smaller A sample is approximately one third the number in the smaller OD sample, but the bundles from A appear in ten times more documents.

5.3 Number of groups of (partially) overlapping or similar lexical bundles (*LB groups*), identified manually

	OD	ODext	A	Aext	AR	ARext
LB groups	29	29	14	16	16	13

Tab. 4.

Similarly, to point 2, the distribution of groups within corpora remains more or less the same, with 29 groups in OD and 13, 14 or 16 groups in both A and AR. Again, this could indicate a bigger "diversity" of formulaic sequence types in the OD corpus, but the appearance might also be related to different average document lengths in the three corpora, as mentioned in section 5.1.

5.4 Distribution of bundle groups (*dist bnd grps*)

Described as x(y), where y is the number of 5-word bundles within the group and x is the number of groups. Examples of frequent longer sequences comprised of partially overlapping five-word bundles in each of the six samples can be seen below the respective numbers of bundle groups, with the first (most frequent) five-word bundle marked in bold. The number of groups in each sample needs to be completed with the number of wording variations in Tab. 5.

	OD	ODext	A	Aext	AR	ARext
dist bnd grps	2(8)	1(8)	x	x	1(8)	1(8)
	2(7)	x	1(10)	x	x	x
	1(6)	2(6)	x	x	x	x

	OD	ODext	A	Aext	AR	ARext
	2(5)	4(5)	1(5)	1(5)	x	x
	6(4)	3(4)	2(4)	x	x	x
	3(3)	4(3)	x	x	3(3)	1(3)
	3(2)	6(2)	3(2)	3(2)	4(2)	4(2)
	7(1)	9(1)	4(1)	9(1)	8(1)	7(1)

Tab. 5.

The almost 30 groups in both OD samples are distributed relatively homogeneously with 1 to even 6 occurrences of bundles consisting of eight, seven, six etc. overlapping 5-word bundles. In samples from the AR corpus, we can notice appearance of groups consisting of 8 bundles. But apart from that, groups of 5 to 2 bundles appear only 1 to 4 times, leading to the conclusion that the distribution in the middle group size range is more heterogeneous here. Bundles consisting of just 5 words, however, appear with similar frequency in most samples, with the exception of the smaller A sample. This observation would suggest that bundles in OD corpus are both more diverse and that they appear with equal regularity, while there is a lower bundle diversity in the A and AR corpora, where the bundles, in turn, constitute more substantial groups (longer sequences).

5.5 Number of similar bundles within a group (*incl word var*)

Described as x(y/w-z-...), where y represents the number of bundles within the group, w, z… are the numbers of bundles containing word variations in the group and x is the number of occurrences for y. Examples of wording variation found in all six samples can be seen below the numerical variation schemes.

	OD	ODext	A	Aext	AR	ARext
incl word var	1(8/3-4-1)	x	1(12/1-7-1-3)	1(12/1-4-1-3-1-2)	x	x
	1(4/2-2)		1(7/2-2-1-2)	1(8/2-2-3-1)		
	1(2/1-1)		1(2/1-1)	1(2/1-1)		

Tab. 5. cont.

This is a subsection of groups identified in 5.4, and here one observes a similarity between the number of wording variations found in both OD and AR corpora. Both A samples contained several groups with more variations and wording similarities. This might be related to the presence of less diverse, albeit longer, bundle sequences especially in the A/Aext corpus, as described in point 5.4.

5.6 Maximal, minimal, average, median and mode values for normalized frequency of individual bundles (*ipm*)

	OD	ODext	A	Aext	AR	ARext
max **ipm**	286.5	274.9	843.0	670.1	434.6	290.0
min ipm	106.7	86	57.8	42.3	33.8	26.8
avr ipm	115.7	166.7	179.1	166.8	100.5	81.2
med ipm	139.7	160.8	100.7	116.2	45.5	40.6
mod ipm	132.7	160.8	127.1	129.9	33.8	35.2

Tab. 6.

Maximal and minimal normalized frequencies remain relatively similar in smaller and larger samples within all three corpora, while the maximum value in OD samples is approximately four times lower than that in A samples, and almost the same/one third lower than in the AR samples. On the contrary, minimal values in OD samples are higher than in both A (two-) and AR corpora (three times). This could be perceived as another trace of distributional structures within bundle groups as described in point 5.4, with more homogeneous distribution of various and more numerous bundle groups in OD samples. However, average ipm values in all three corpora are relatively similar, while there is a bigger similarity between median and mode values in OD and A corpora than in the AR corpus.

5.7 Maximal, minimal, average, median and mode values of bundle dispersion per sample ratio (*dis/s*)

Calculated as dispersion value divided by the number of all documents within a sample. Dispersion (*dis*) is calculated as the number of first-time appearances of a given bundle in sample documents, that is, as the number of documents in the a sample where the bundle appears at least once.

	OD	ODext	A	Aext	AR	ARext
max **dis/s**	50,3	53,8	60,1	55,7	83,6	65,9
min dis/s	6,7	8,5	2,8	0,04	0,9	0,9
avr dis/s	24,7	29,3	21,7	26,2	28,3	31,2
med dis/s	25,1	32,3	16,1	21,2	20	32,2
mod dis/s	24,5	32,4	4,9	55,6	1,8	0,9

Tab. 7.

In this respect, individual bundles appear to the maximum value in around 50 percent of OD sample documents, whereas the maximum values for normalized maximal dispersion are slightly higher (50–60 percent) in A samples, and even higher (more than 60/80 percent) in AR samples. Minimal values seem to behave in an inverted order, with the highest minimal value in OD and lowest in AR samples.

Both average and median values for the normalized dispersion, however, are relatively similar in all six samples. The mode values decrease in smaller OD, A and AR samples, respectively, while this value is relatively higher in larger OD and A samples, but below 1 in the larger AR sample.

5.8 Communicative functions of the bundles

Apart from distributional and proportional characteristics summarized in the table above, the 6 overview sample tables enabled to compare bundles in three corpora according to communicative (discourse) functions of bundles. The question would be, how to compare communicative functions of bundles in different genres. Here, it is possible to use a common classification scheme of these functions, as the one originally used by D. Biber et al. (stance bundles, discourse organizers, referential bundles) [20]. It is also possible to group the present bundles in each genre according to the more or less schematic textual structure of administrative texts, such as decisions, acts, reports.

5.8.1 Communicative functions of the bundles found in judicial decisions

Communicative functions of bundles found in judicial decisions can be described by placing the individual bundle into four categories, corresponding to basic parts of a decision, where these bundles normally appear. These parts are

a) heading part which identifies the case, the involved parties and other circumstances;
b) enunciation/verdict part which pronounces the verdict, often in several counts;
c) reasoning part which presents the reasons that have led the court to reach its verdict;
d) instruction part which advises the party suffering as a result of the decision on possible remedies.

Following this textual structure, the lexical bundle groups found in both samples of OD corpus fall in line with distributional characteristics presented in point 5.4 mostly in instruction and reasoning parts.

Communicative function/Type	LB groups in OD sample	Examples	ipm
instruction part	15	1. *len do uplynutia lehoty na (podanie odvolania)* 'only until expiration of (appellate) period'	286
		2. *z akých dôvodov sa rozhodnutie (považuje za nesprávne)* '(what are the) grounds for (considering) the decision (as incorrect)'	272
		3. *v akom rozsahu sa (rozhodnutie) napáda* '(what is) the extent of appeal'	269

Communicative function/Type	LB groups in OD sample	Examples	ipm
reasoning part	10	1. *(o) súdnych exekútoroch a exekučnej činnosti* 'on judicial enforcers and on enforcement proceeding'	241
		2. *a náhrada za stratu času* 'and compensation of lost time'	206
		3. *v konkurze alebo splátkovým kalendárom* 'bankrupt or by repayment plan'	173
verdict part	3	1. *nárok na náhradu trov konania* 'entitlement to covering of costs of enforcement proceeding'	186
		2. *odo dňa doručenia platobného rozkazu* 'after the service of the charging order in writing'	168
		3. *(ktoré) môžu byť uspokojené iba v* '(which) can only be met in'	122
heading part	0		
other	1	*MENE SLOVENSKEJ REPUBLIKY Okresný súd* 'BEHALF OF THE SLOVAK REPUBLIC District court'	121
Communicative function/Type	LB groups in ODext sample	Examples	ipm
instruction part	19	1. *a čoho sa odvolateľ domáha* 'what is pursued by the appellant'	266
		2. *alebo postup súdu považuje za (nesprávny)* 'or considers court's procedural measures (to be unlawful)'	259
		3. *v čom sa toto rozhodnutie* '(what are the grounds for considering) this decision'	253
reasoning part	7	1. *(o) udelenie poverenia na vykonanie exekúcie* '(for) granting of authorization for enforcement'	274
		2. *právo na náhradu trov konania* 'entitlement to covering of costs of enforcement proceeding'	205
		3. *(o) súdnych exekútoroch a exekučnej činnosti* 'on judicial enforcers and on enforcement proceeding'	142
verdict part	2	1. *nemá právo na náhradu trov* 'is not entitled to covering of costs of enforcement proceeding'	134
		2. *(s) úrokom z omeškania vo výške* '(with) late charges of'	107
heading part	0		
other	1	*MENE SLOVENSKEJ REPUBLIKY Okresný súd* 'BEHALF OF THE SLOVAK REPUBLIC District court'	97

Tab. 8.

5.8.2 Communicative functions of the bundles found in acts

The 5-word bundle groups found in acts are mostly parts of formulas, above all the amendment and supplementing formula, consisting of twelve 5-word bundles (including variations); promulgation formula, through which an act comes into being, so to speak; or formulas referring to the body of legal text itself, signalling a footnote or a change to the text.

Communicative function/Type	LB groups in A/ Aext sample	Examples	ipm
amendment and supplementing formula	3/3	1. *a o zmene a doplnení (niektorých zákonov)* 'and on amendment and supplementing of (several statutes)'	843/670
		2. *sa mení a dopĺňa takto* 'is amended and supplemented in the following way'	316/241
		3. *ktorým sa mení a dopĺňa (zákon)* 'by which (the statute) is amended and supplemented'	106/123
text reference formula	2/3	1. *Poznámka pod čiarou k odkazu* 'Footnote with respect to reference'	650/528
		2. *Poznámky pod čiarou k odkazu* 'Footnotes with respect to reference'	198/174
		3. *sa na konci pripájajú tieto (slová)* 'following (words) are supplemented to the end'	172/149
promulgation formula	1/1	*(Národná rada Slovenskej republiky) sa uzniesla na tomto zákone* '(The National Council of the Slovak Republic) has enacted the following statute'	127/130
other	8/9	1. *nálezu Ústavného súdu Slovenskej republiky* '(of the) finding of the Constitutional Court of the Slovak Republic'	145/114
		2. *(sociálnoprávnej) ochrany detí a sociálnej kurately* 'of the child welfare services'	89 (A)
		3. *ak tento zákon neustanovuje inak* 'unless this Act stipulates otherwise'	76 (Aext)

Tab. 9.

5.8.3 Communicative functions of the bundles found in annual reports

The most common formula (amendment and supplementing f.) found in acts is also that which is most commonly found in annual reports (here, it consist of 8 shorter bundles), while, apart from that, bundles in the annual reports consist to some extent of proper names referred to in the report texts.

Communicative function/Type	LB groups in AR/ARext sample	Examples	ipm
amendment and supplementing formula	2/2	1. *a o zmene a doplnení (niektorých zákonov)* 'and on amendment and supplementing of (several statutes)'	434/290
		2. *ktorým sa mení a dopĺňa (zákon)* 'by which (statute) is amended and supplemented'	93/69
proper name	3/2	1. *Ministerstva pôdohospodárstva a rozvoja vidieka (Slovenskej republiky)* '(of the) Ministry of Agriculture and Rural Development (of the Slovak Republic)'	50 (AR)
		2. *Štátna vedecká knižnica v Prešove* 'State Science Library in Prešov'	40/27
		3. *Ministerstva životného prostredia Slovenskej republiky* '(of the) Ministry of Environment of the Slovak Republic'	35/26

Communicative function/Type	LB groups in AR/ARext sample	Examples	ipm
other	11/9	1. *(pri) výkone práce vo verejnom záujme* 'during the public interest service'	66/63
		2. *škôl a školských zariadení v* '(of) schools and school facilities in'	59/46
		3. *štátnej správy starostlivosti o životné (prostredie)* '(of) national environmental administration'	43/31

Tab. 10.

6 CONCLUSION

The aim in presenting the quantitative characteristics was not to prove that judicial decisions are simply more or less formulaic (that is, schematic or prefabricated) than acts or annual reports. Every text, register, style or genre has its own means by which it can be considered to accomplish its communicative function, even when these methods can become subject to dispute, because they can be perceived as not sufficiently effective or stylistically balanced, as F. Štícha points out with reference to the style of judicial decisions [26, pp. 71–72]. There is no "linear", deterministic connection between the use of certain lexis in a given text and its communicative function. Nor can we afford to neglect the existence of "style-mixes" and "transitional areas" ([27], cited in [4, p. 208]). This is especially true in the case of judicial decisions, which often contain the explicit language given in testimonies, as well as specialized legal terms and analytic multi-word expressions. This opens for the possibility that the feature called formulaicity, as well as complex linguistic phenomena in general, might be productively studied by applying methods of multidimensional analysis, such as this has been showcased in [24].

Nevertheless, the lexical bundles found in samples of Slovak judicial decisions are word sequences that constitute the textual result of an institutionally regulated social interaction between these institutions (courts) and involved parties, be it physical and/or juridical personae. The decisions are thus directed both at the realms of "normativity" [28, pp. 84–113] and factuality, as the decisions represent a multifaceted, possibly complex [29, p. 216] process of law application (the realm of regulations and norms) regarding the factual case (involving individuals or legal entities). As shown in Section 5, point 9, a noticeable quantity of bundles found in Slovak judicial decisions informs involved parties when deciding how to react to the implications of a court ruling (e. g. on how to appeal) and upon what rationale has the court based its verdict. The intersection of individuals and individual cases (factuality) and regulations (normativity) might, then, require a set of detailed regulations (mostly sections 363, 364 and 365 of the Slovak Civil Contentious Procedure Code), that constitute the reasoning and instructional components of its

decisions. This, in turn, may have resulted in the relative high diversity and relative low variation of bundles in Slovak judicial decisions, as compared to acts.

The acts, on the other hand, address not so much particular individuals (even though they are produced in institutional settings of legislative bodies) as much as decisions arrived at in courts. The acts inform the broader public regarding how social interaction in various domains should play out, while also outlining the implications (sanctions) for non-compliance with these regulations. This means that acts express both norms and a model-like representation of reality, while operating more in the realm of norms and ideals, in that they both incorporate new regulations into the body of legislation and change existing regulations. Lexical bundles found in the act samples bear witness of these legislative procedures, as they consist mainly of traditional formulae used when referring to the actual and/or amended wording of acts, or formulas for the promulgation of new regulations. The aforementioned reference formulae are also found as bundles in Slovak annual reports, while there are also bundles denoting proper names of institutions or document titles. These bundles constitute (perhaps because they are more bound to specific report subjects) shorter bundle groups. Some future research into patterns emerging via formulaic sequences in texts produced in institutional settings might constitute a promising field of study.

ACKNOWLEDGEMENTS

The paper has been written within the Slovak National Corpus project supported by the Slovak Academy of Sciences, Ministry of Education, Science, Research and Sport of the Slovak Republic, Ministry of Culture of the Slovak Republic and the Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences.

References

[1] Prusák, J. (2001). Teória práva. Bratislava: Vydavateľské oddelenie Právnickej fakulty UK 2001, 331 p.

[2] Engberg, J. (2013). Legal linguistics as a mutual arena for cooperation. Recent developments in the field of applied linguistics and law. AILA Review 26, pages 24–41.

[3] Holländer, P. (2014). Interpretácia práva. Justičná akadémia. Available at: https://www.ja-sr.sk/files/Interpretacia_prava.pdf.

[4] Imrichová, M. (2014). Jazyk právnych textov so zreteľom na špecifiká právnej lexiky. In J. Kesselová, M. Imrichová and M. Ološtiak (eds.), Registre jazyka a jazykovedy (I). Prešov. Filozofická fakulta Prešovskej univerzity v Prešove, pages 207–212.

[5] Gahér, F., Števček, M., and Braxatoris, M. (2019). Nástroje a pravidlá produkcie a interpretácie koncízneho textu (s osobitným zreteľom na normativitu). Jazykovedný časopis, 70(1), pages 73–94.

[6] Berūkštienė, D. (2017). A corpus-driven analysis of structural types of lexical bundles in court judgments in English and their translation into Lithuanian. Kalbotyra 70, pages 7–31.

[7] Kačala, J. (2010). Zložené útvary v jazyku. Martin: Vydavateľstvo Matice slovenskej. 134 p.

[8] I. MacKenzie and M. A. Kayman (eds.). (2018). Formulaicity and Creativity in Language and Literature. Routledge. 126 p.

[9] Kováříková, D. (2017). Kvantitativní charakteristiky termínů. Praha: Nakladatelství Lidové noviny. 136 p.
[10] Chlumská, L. (2017). Překladová čeština a její charakteristiky. Praha: Nakladatelství Lidové noviny. 150 p.
[11] Grabowski, Ł., and Forsyth, R. S. (2015). Is there a formula for formulaic language? Poznan Studies in Contemporary Linguistics 51(4), pages 511–549.
[12] Wray, A., and Perkins, M. (2000). The functions of formulaic language: An integrated model. Language & Communication 20(1), pages 1–28.
[13] Biber, D., Johansson S., Leech G., Conrad S., and Finegan E. (1999). The Longman Grammar of Spoken and Written English. Harlow (Essex): Longman, 1204 p.
[14] Biber, D., and Barbieri, F. (2007). Lexical bundles in university spoken and written registers. English for Specific Purposes 26, pages 263–286.
[15] Grabowski, Ł. (2014). On lexical bundles in Polish patient information leaflets: a corpus-driven analysis. Studies in Polish Linguistics 9(1), pages 21–43.
[16] Breeze, R. (2013). Lexical bundles across four legal genres. International Journal of Corpus Linguistics 18(2), pages 229–253.
[17] Biel, Ł. (2017). Lexical bundles in EU law: the impact of translation process on the patterning of legal language. In Phraseology in legal and institutional settings. A corpus-based interdisciplinary perspective (e-book). S. Goźdź-Roszkowski and G. Pontrandolfo (eds.). London, New York: Routledge, pages 10–26.
[18] Goźdź-Roszkowski, S. (2018). Between corpus-based and corpus-driven approaches to textual recurrence: Exploring semantic sequences in judicial discourse. In J. Kopaczyk and J. Tyrkkö (eds.), Applications of Pattern-driven Methods in Corpus Linguistics, pages 131–158.
[19] Hyland, K. (2008). As can be seen: lexical bundles and disciplinary variation. English for Specific Purposes 27, pages 4–21.
[20] Biber, D., Conrad, S., and Cortes, V. (2004). If you look at...: lexical bundles in university teaching and textbooks. Applied Linguistics 25 (3), pages 371–405.
[21] Wiechmann, D., and Kerz, E. (2016). Formulaicity as a determinant of processing efficiency: investigating clause ordering in complex sentences. English Language and Linguistics 20(3), pages 421–437.
[22] Tománková, V. (2016). Lexical bundles in legal texts corpora – selection, classification and pedagogical implications. Discourse and interaction 9(2), pages 75–94.
[23] Grabowski, Ł. (2018). Fine-tuning lexical bundles: A methodological reflection in the context of describing drug-drug interactions. In J. Kopaczyk and J. Tyrkkö (eds.), Applications of Pattern-driven Methods in Corpus Linguistics, pages 15–56.
[24] Cvrček, V., Laubeová, Z., Lukeš, D., Poukarová, P., Řehořková, A., and Zasina, A. J. (2020). Registry v češtině. Praha: Nakladatelství Lidové noviny. 234 p.
[25] Čech, R., Garabík, R., and Altmann, G. (2015). Testing the Thematic Concentration of Text, Journal of Quantitative Linguistics, 22(3), pages 215–232.
[26] Štícha, F. (1985). O jazyce soudních rozhodnutí. Naše řeč, 68(2), pages 68–77.
[27] Slančová, D. (1999). Reč autority a lásky: reč učiteľky materskej školy orientovaná na dieťa. Opis registra. Prešov: Filozofická fakulta Prešovskej univerzity. 224 p.
[28] Procházka, R., and Káčer, M. (2019). Teória práva. 2. vydanie. Bratislava: C. H. Beck. 304 p.
[29] Gábriš, T. (2018). Sudcovské rozhodovanie z pohľadu súčasného právneho realizmu. In T. Gábriš (ed.), Sudcovské rozhodovanie. Záruky a prekážky spravodlivého procesu. Bratislava: Wolters Kluwer, pages 216–233.

POKYNY PRE AUTOROV

Redakcia JAZYKOVEDNÉHO ČASOPISU uverejňuje príspevky **bez poplatku** za publikovanie.
Akceptované jazyky: všetky slovanské jazyky, angličtina, nemčina. Súčasťou vedeckej štúdie a odborného príspevku je abstrakt v angličtine (100 – 200 slov) a zoznam kľúčových slov v angličtine (3 – 8 slov).
Súčasťou vedeckej štúdie a odborného príspevku v inom ako slovenskom alebo českom jazyku je zhrnutie v slovenčine (400 – 600 slov) – preklad do slovenčiny zabezpečí redakcia.
Posudzovanie príspevkov: vedecké príspevky sú posudzované anonymne dvoma posudzovateľmi, ostatné príspevky jedným posudzovateľom. Autori dostávajú znenie posudkov bez mena posudzovateľa.
Technické a formálne zásady:

- Príspevky musia byť v elektronickej podobe (textový editor Microsoft Word, font Times New Roman, veľkosť písma 12 a riadkovanie 1,5). V prípade, že sa v texte vyskytujú zvláštne znaky, tabuľky, grafy a pod., je potrebné odovzdať príspevok aj vo verzii pdf alebo vytlačený.
- Pri mene a priezvisku autora je potrebné uviesť pracovisko.
- Text príspevku má byť zarovnaný len z ľavej strany, slová na konci riadku sa nerozdeľujú, tvrdý koniec riadku sa používa len na konci odseku.
- Odseky sa začínajú zarážkou.
- Kurzíva sa spravidla používa pri názvoch prác a pri uvádzaní príkladov.
- Polotučné písmo sa spravidla používa pri podnadpisoch a kľúčových pojmoch.
- Na literatúru sa v texte odkazuje priezviskom autora, rokom vydania a číslom strany (Horecký, 1956, s. 95).
- Zoznam použitej literatúry sa uvádza na konci príspevku (nie v poznámkovom aparáte) v abecednom poradí. Ak obsahuje viac položiek jedného autora, tie sa radia chronologicky.

Bibliografické odkazy:

- knižná publikácia: ONDREJOVIČ, Slavomír: Jazyk, veda o jazyku, societa. Bratislava: Veda, vydavateľstvo SAV 2008. 204 s.
- slovník: JAROŠOVÁ, Alexandra – BUZÁSSYOVÁ, Klára (eds.): Slovník súčasného slovenského jazyka. H – L. [2. zv.]. Bratislava: Veda, vydavateľstvo SAV 2011. 1088 s.
- štúdia v zborníku: ĎUROVIČ, Ľubomír: Jazyk mesta a spisovné jazyky Slovákov. In: Sociolinguistica Slovaca 5. Mesto a jeho jazyk. Ed. S. Ondrejovič. Bratislava: Veda, vydavateľstvo SAV 2000, s. 111 – 117.
- štúdia v časopise: DOLNÍK, Juraj: Reálne vz. ideálne a spisovný jazyk. In: Jazykovedný časopis, 2009, roč. 60, č. 1, s. 3 – 12.
- internetový zdroj: Slovenský národný korpus. Verzia prim-5.0-public.all. Bratislava: Jazykovedný ústav Ľudovíta Štúra SAV 2010. Dostupný na: http://korpus.juls.savba.sk [cit. DD. MM. RRRR].

INSTRUCTION FOR AUTHORS

JOURNAL OF LINGUISTICS publishes articles **free of publication charges**.
Accepted languages: all Slavic languages, English, German. Scientific submissions should include a 100-200 word abstract in English and a list of key words in English (3-8 words).
Scientific articles in a language other than Slovak or Czech should contain a summary in Slovak (400-600 words) – translation into Slovak will be provided by the editor.
Reviewing process: scientific articles undergo a double-blind peer-review process and are reviewed by two reviewers, other articles by one reviewer. The authors are provided with the reviews without the name of the reviewer.
Technical and formal directions:

- Articles must be submitted in an electronic form (text editor Microsoft Word, 12-point Times New Roman font, and 1.5 line spacing). If the text contains special symbols, tables, diagrams, pictures etc. it is also necessary to submit a pdf or printed version.
- Contributions should contain the full name of the author(s), as well as his/her institutional affiliation(s).
- The text of the contribution should be flush left; words at the end of a line are not hyphenated; a hard return is used only at the end of a paragraph.
- Paragraphs should be indented.
- Italics is usually used for titles of works and for linguistic examples.
- Boldface is usually used for subtitles and key terms.
- References in the text (in parentheses) contain the surname of the author, the year of publication and the number(s) of the page(s): (Horecký, 1956, s. 95).
- The list of references is placed at the end of the text (not in the notes) in alphabetical order. If there are several works by the same author, they are listed chronologically.

References:

- Monograph: ONDREJOVIČ, Slavomír: Jazyk, veda o jazyku, societa. Bratislava: Veda, vydavateľstvo SAV 2008. 204 p.
- Dictionary: JAROŠOVÁ, Alexandra – BUZÁSSYOVÁ, Klára (eds.): Slovník súčasného slovenského jazyka. H – L. [2. zv.]. Bratislava: Veda, vydavateľstvo SAV 2011. 1088 p.
- Article in a collection: ĎUROVIČ, Ľubomír: Jazyk mesta a spisovné jazyky Slovákov. In: Sociolinguistica Slovaca 5. Mesto a jeho jazyk. Ed. S. Ondrejovič. Bratislava: Veda, vydavateľstvo SAV 2000, pp. 111 – 117.
- Article in a journal: DOLNÍK, Juraj: Reálne vz. ideálne a spisovný jazyk. In: Jazykovedný časopis, 2009, Vol. 60, No 1, pp. 3 – 12.
- Internet source: Slovenský národný korpus. Verzia prim-5.0-public.all. Bratislava: Jazykovedný ústav Ľudovíta Štúra SAV 2010. Dostupný na: http://korpus.juls.savba.sk [cit. DD. MM. YEAR].

ISSN 0021-5597 (tlačená verzia/print) ISSN 1338-4287 (verzia online) MIČ 49263

JAZYKOVEDNÝ ČASOPIS

VEDECKÝ ČASOPIS PRE OTÁZKY TEÓRIE JAZYKA

JOURNAL OF LINGUISTICS

SCIENTIFIC JOURNAL FOR THE THEORY OF LANGUAGE

Objednávky a predplatné prijíma/Orders and subscriptions are processed by:
SAP – Slovak Academic Press, s. r. o., Bazová 2, 821 08 Bratislava
e-mail: sap@sappress.sk

Registračné číslo 7044

Evidenčné číslo 3697/09

IČO vydavateľa 00 167 088

Ročné predplatné pre Slovensko/Annual subscription for Slovakia: 12 €, jednotlivé číslo 4 €
Časopis je v predaji v kníhkupectve Veda, Štefánikova 3, 811 06 Bratislava 1